The Dissertation Committee for Kam Hon Hoi Certifies that this is the approved version of the following dissertation:

Global survey of the immunoglobulin repertoire using next generation sequencing technology

Committee:

George Georgiou, Supervisor

Gregory Ippolito

Ning (Jenny) Jiang

Edward Marcotte

Haley Tucker

Global survey of the immunoglobulin repertoire using next generation sequencing technology

by

Kam Hon Hoi, B.S.; B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2014

Dedication

This dissertation is dedicated to my parents and my grandparents for their life-long support and guidance. Their optimism and joy for life has inspired me to appreciate life with a similar outlook.

Acknowledgements

I would like to express my sincere appreciation and gratitude to my adviser Dr. George Georgiou. I thank him for not only his wit and charm but also his attitude and concern for his friends and students, I will always be inspired to become a role model like him. And, I valued his advice and vision in the building of my scientific career. Once again, I would like to sincerely thank him for everything that he has taught me.

I would also like to thank two other important mentors during my graduate education --- Dr. Sai Reddy and Dr. Gregory Ippolito. They opened the door to immunology for me. Not only did they encourage me to take the first step, but they also helped to guide me along the way. I greatly appreciate everything they have taught me to survive the trail. I can say that the hike in immunology has not been easy, but the scenery certainly was worth the effort. I greatly appreciate everything that they have taught me in preparation for embarking down the next trail in immunology.

As for graduate education, I think of it like a marathon since it is a test of perseverance and endurance. It would have been sad to run it all alone but I was fortunate to have great running mates. They not only shared my hardships but more importantly shared my happiness as well. I would like to express my greatest gratitude to Dr. Constantine Chrysostomou and Dr. Erik Quandt who were always the first to pick me up when I fell and the first to cheer even when I ran just one millisecond faster. Also, everyone that kept the run fun and lively, especially, Dr. Jason Lavinder, Dr. William Kelton, Dr. Yariv Wine, Giulia Agnello, Scott Kerr, Ellen Murrin, Dr. Sebastian Schätzle, Dr. Moses Donkor, Erik Johnson, Wissam Charab, Jiwon Lee, Dr. Tae Hyun Kang, and Elizabeth Miller. I am grateful to be running with such a lovely bunch.

Global survey of the immunoglobulin repertoire using next generation sequencing technology

Kam Hon Hoi, Ph.D.

The University of Texas at Austin, 2014

Supervisor: George Georgiou

Specific and sensitive recognition of foreign agents is a critical attribute of the overall effective immune system required for maintaining host protection against challenge from pathogenic cells. In the humoral arm of the immune system, this recognition attribute is carried out by the cell surface bound immunoglobulin-like receptors (BCR) and its soluble forms i.e. antibodies. Over several million years of evolution, the immune system has adopted several strategies for diversifying the antibody sequence and thus its ability to recognize an astronomical variety of molecules through the combinatorial assembly of a small number of DNA segments or genes. Among these immunoglobulin gene diversification strategies, antibody somatic VDJ recombination and junctional diversity are the fundamental mechanisms in generating a broad range of antibody specificities. Understanding how the genetic diversity of antibodies is affected in health and disease is critical for a wide range of medical applications, from vaccine evaluation to diagnostics and therapeutics discovery. Because of the very large number of distinct antibodies encoded by the more than 10 billion B cells in humans, it is essential to use high throughput next generation sequencing technologies in order to obtain an adequate sampling of the sequences and relative abundance of different antibodies expressed by B cells in clinical samples. The process requires rigorous methods for first,

experimentally determining the sequences of antibodies in a sample and for second, informatics tools designed for distilling this information for practical purposes. This dissertation describes a variety of experimental approaches and informatics tools developed for the determination and mining of the antibody repertoire. The information from this work has not only led to major conclusions in efficient antibody discovery but also regarding the nature of the antibody repertoire in healthy individuals, in rabbits, in volunteers following vaccination, and in HIV-1 patients.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The well-being of a multi-cellular organism is dependent upon the efficacy of its protection strategies against foreign pathogens and pathological agents. Such protection in higher-order organisms is orchestrated by the innate and the adaptive immune responses. Collectively regarded as the immune system, these protection strategies are undeniably one of the most complex molecular and cellular networks in maintaining health. Many medical interventions and remedies entail the restoration or the exploitation of the various facets of the immune system to sustain health [1], [2]. The immune system can be broadly categorized into four unique functionalities: immunological recognition, immune effector functions, immune regulation, and immunological memory [1]. This dissertation is focused on the use of next generation DNA sequencing technology to survey the antibody repertoire pertaining to immunological recognition. First, however, a brief overview of the major components of the immune system and their respective functions are included herein to help illustrate the significance of the studies presented in this dissertation.

THERAPEUTIC SIGNIFICANCE OF ANTIBODY REPERTOIRE STUDIES AND ANTIBODY DISCOVERY TECHNOLOGIES

Manipulation of the immune system has long historical records and has fortified its place in medicine. Pioneered by Edward Jenner in the late 18th century, smallpox vaccination was the first vaccination that made a significant impact on the greatest medical woe of that era and eventually led to the eradication of the smallpox virus [2]. A century later, Emil Adolf von Behring and Paul Ehrlich made a medical breakthrough using antiserum therapy to combat diphtheria and tetanus. As a result, both scientists

were awarded with Nobel Prize in the early 20th century [2]. In both medical breakthroughs, antibodies played an irreplaceable role in the success of the therapies (overview of antibody is provided in the following section). Throughout the ages, there have been unwavering efforts to develop monoclonal antibodies (mAbs) for the treatment of diseases such as cancers (leukemia, lymphoma, breast cancer, etc.), autoimmune diseases (Systemic Lupus Erythematosus, Rheumatoid Arthritis, etc.), viral infections (Influenza, HIV, etc.), Graft-versus-host diseases, and allergies. In fact, the significance of mAbs in tackling medical needs can be reflected in its expanding global market share. Monoclonal antibodies have maintained their leading role in the biotechnology industry for the past few decades [3], [4]. And, the worldwide market for mAbs has grown from $6.9 billion in 2003 to about $24.8 billion in 2007 with an annual growth rate estimated at about 38% [3]. In the US alone, mAbs have reached about $24.6 billion in sales reaping approximately 18.3% growth over the prior year [4]. The 2012 sales in mAbs is equivalent to the sales made by the next two classes of drugs combined, namely hormones and growth factors.

One of the reasons for the huge success of mAbs is the proven clinical efficacy highlighted by several commercial blockbuster mAbs such as Rituxan (rituximab), Humira (adalimumab), and Avastin (bevacizumab). With advancement in the humanizing antibodies and Fc glycoengineering, the side-effects of mAbs can be further minimized. Another reason mAbs are the "weapon" of choice for combating diseases is their high specificity, hence, reducing undesired off-targets effects. For example, the trait of high specificity is also capitalized on the targeted delivery of chemotherapy. Antibody-drug conjugates such as Kadcyla (trastuzumab emtansine) is Herceptin (trastuzumab) conjugated to cytotoxic agent mertansine (DM1) where it can be used to specifically

deliver the toxic agent to HER2+ breast cancer cells [5]. As importantly, the Fc fragment (constant region of the antibody) can modulate different immune effector functions (discussed in subsequent section). And, this particular trait has also been capitalized to improve existing mAbs; for example, Gazyva (obinutuzumab) is the first mAbs approved by the Food and Drug Administration (FDA) through breakthrough therapy designation (BTD) and it is also the first approved glycoengineered and Fc-engineered antibody. The glycoengineered Fc portion of Gazyva has been shown to improve direct killing via antibody-dependent cell-mediated cytotoxicity (ADCC) by enhancing the engagement of the Fc receptors on effector cells [6]. Because antibodies possess such therapeutic benefits, gaining knowledge on the antibody repertoire during immune responses can help with the discovery of therapeutically relevant antibodies. Since antibody plays such a central role in this dissertation, it is important to review this molecule in more details in the next section.

ANTIBODIES ARE IMPORTANT PROTEIN MEDIATORS OF EFFECTOR FUNCTIONS AND IMMUNOLOGICAL RECOGNITION

The antibody molecule in its native form is a homodimer made of two heterodimeric units. Each heterodimeric unit consists of two different polypeptide chains: a light chain (shown in red and yellow regions in Figure 1) and a heavy chain (shown in blue and green regions in Figure 1) [7], [8]. Two identical heavy-light chain heterodimeric units are then bound together by disulfide bonds to form the full-length antibody. Considering just the heavy-light chain heterodimeric unit, the light chain contains two domains: a variable (VL) region (red regions in Figure 1) and a constant (CL) region (yellow regions in Figure 1). The heavy chain of an IgG isotype antibody contains four domains: a variable (VH) region (blue regions in Figure 1) and three

constant (CH1, CH2, CH3) regions (green regions in Figure 1). The different isotypes represent the different conserved constant region sequences encoded in the host's genome and it will be discussed in the later paragraphs within this section.

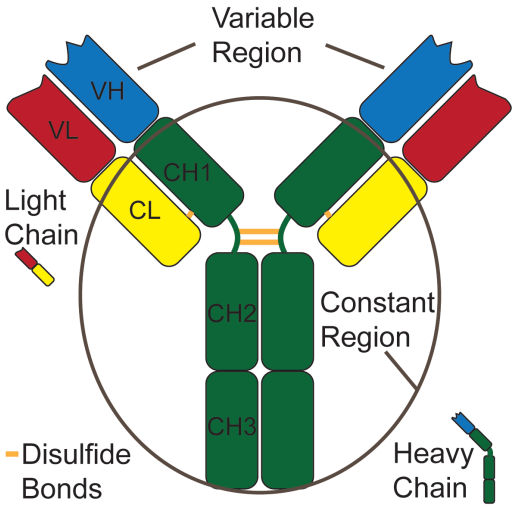

Figure 1: Schematic diagram of a full-length Immunoglobulin G (IgG)

Structurally, each region adopts a general tertiary organization folded by antiparallel β-sheets stabilized through disulfide bonds and a hydrophobic core. In terms of the modular domains, the arm consisting of the VL, VH, CL, and CH1 regions are known as the Fab fragment (50 kDa) and the stem consisting of the dimeric form of CH2 and CH3 regions are known as the Fc fragment (50 kDa). Hence, a full-length IgG antibody molecule is about 150 kDa. The different fragments can be generated using the protease papain digestion to break an antibody into two Fab fragments and one Fc fragment.

As mentioned before, the human genome encodes different conserved groups of antibody constant region sequences known as isotypes, each of which are associated with different immune response functions. In most mammals, there are generally five different isotypes: IgM, IgG, IgA, IgE, and IgD. The antibody most commonly found in the serum is IgG and its structure is one of the first to have been extensively studied and thus often referenced as the example of an antibody. The various antibody isotypes differ in their sequence composition in the CH1-CH3 domains. Additionally, IgM and IgE isotype antibodies contain an extra CH4 domain. Apart from the sequence composition of the Fc fragments, IgM and IgA form multimers as depicted in Figure 2.

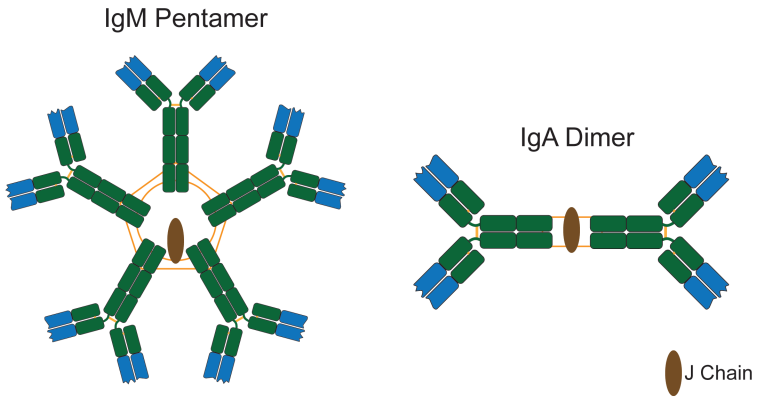

Figure 2: Schematic diagram of multimeric form of Immunoglobulin M (IgM) and Immunoglobulin A (IgA)

Multimerization of IgM and IgA requires a small tail of 18 amino acids containing a cysteine residue that is critical for multimer formation. A separate polypeptide, the J chain (15 kDa), mediates multimer formation by linking the tail's cysteine of each unit together [9], [10]. Multimer formation was found to have significant implication in enhancing binding strength against repeating epitopes via avidity effects

(i.e. an increase in the effective concentration of the binding sites) on bacterial pathogens and facilitating IgA transport through epithelia for mucosal secretion [1], [9], [10]. The Fc domain of the antibody fragment also affects the degree to which an antibody can engage various effector immune cells via interactions with their membrane-bound Fc receptors (FcRs) [11]. The role of FcRs is known to mediate antibody-induced immune response upon binding to the Fc fragment of an antibody. The interactions between the FcRs and the different isotypes are specific; therefore, there are usually as many different classes of FcRs as the isotypes. Depending on the cytoplasmic domain of an FcR, either containing the immunoreceptor tyrosine-based activation motifs (ITAM) or the immunoreceptor tyrosine-based inhibitory motifs (ITIM), an FcR can transduce either activation or inhibitory signals for a cellular response. Hence, effector cells (natural killer cells, T cells, B cells, granulocytes, dendritic cells, and macrophages) expressing different amounts and types of FcRs can respond differently to complexes of pathogens bound by antibodies of various isotypes. Although it is not within the scope of this chapter to further elaborate on the background of Fc modulation, it is noteworthy to at least mention that the Fc fragment can further modulate its interactions with effector cells through the nature of the glycan that is appended to a single Asn residue in the Fc polypeptide. For example, the presence of fucose, galactose, mannose, N-acetylglucosamine, or sialic acid [11], and irregular levels of certain types of glycosylation are observed in patients with various autoimmune diseases [12]. Therefore, the Fc fragment of an antibody possesses important properties that can modulate the degree of immune effector functions.

Without proper immunological recognition, immune effector functions would be modulated non-specifically. The Fab fragment of the antibody molecule is responsible for

immunological recognition. Both the heavy and light variable regions of the Fab fragment form the molecular interface that interacts and recognizes antigens. Interactions between the antigen and the variable regions are mediated by non-covalent forces and may vary in strength depending on the conformational fitting between the variable region and the antigen epitope. Each unique variable region that fits to a unique antigen epitope is called an idiotype. Since idiotype is a function of the structure of the variable regions and structure is dependent upon the amino acid sequence of the antibody polypeptide chain, mechanisms that can generate sequence diversity in antibodies are key to amassing a wide variety of idiotypes. This broad sequence/structure space is often sufficient for surmounting the molecular diversity of possible pathologic molecules. Such a high degree of diversity can only be achieved by recombination and rearrangement of multiple germline variable gene segments (VDJ for V-Heavy chain and VJ for V-Light chain) and the recombinantly joined segments form the naïve antibody repertoire. (A more detailed explanation of both the recombination and the B cell affinity maturation process is discussed in their respective sections.) Briefly, naïve antibodies bind to antigens with low affinity and are then optimized (with respect to affinity) by mutagenesis and a selection process that eventually results in antibodies of exquisite selectivity and affinity. This process is called somatic hypermutation (SHM) and targets predominantly the hypervariable regions of the variable domain that are directly involved in the active binding to the antigens, which can also be observed by antigen-antibody co-crystal structures [13]–[15]. The sequence variability within the hypervariable sections of the Variable Light chain and the Variable Heavy chain regions [16]–[18] in healthy adult peripheral blood B cell antibody sequences is shown in Figure 3. The regions of the variable domain where sequence variability is significantly higher (i.e. the hypervariable regions) relative to the neighboring conserved sequence segments are also called the

complementarity-determining regions (CDR) and they represent the contacting sites with the antigens as mentioned above [14], [15], [19]. Since variability is highly elevated in these sections, the uniqueness of an antibody sequence is usually found within the CDRs. On each of the Variable Light and Variable Heavy chain, the hypervariable regions are denoted as CDR1, CDR2, and CDR3 in Figure 3.

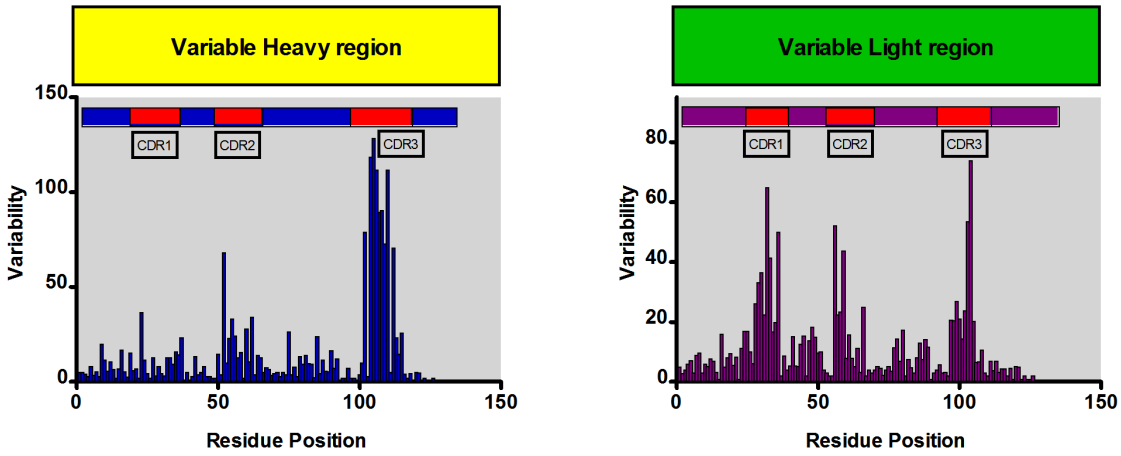

Figure 3: Kabat-Wu variability coefficient plot for Heavy and Light (Lambda) chain variable region

Note: Variability = k/n*N where k is the different numbers of amino acids in the position; n is the number of occurrence for the prevalent amino acids in the position; N is the total number of sequences analyzed

In particular, the CDR3, the section closest to the C terminus on both chains, contains the highest degree of variability and it has the potential to be as unique as a molecular barcode for an antibody sequence. Together, the six CDRs have been shown to be largely responsible for the antigen specificity of the antibodies [7], [8], [14], [15]; hence, knowledge of the sequence composition in the CDRs and its relationship to conformational structure is essential in immunological studies. The degree of recognition

is further refined with a second level of diversity via the combinatorial pairings of the Variable Light chain and the Variable Heavy chain [20]. Therefore, it is just as critical to be able to identify the native pairing of the Variable Heavy and the Variable Light chains in elucidating binding specificity and affinity. There are numerous technical advances that focus on identifying native VH/VL pairs and the topic is discussed in greater detail in a technical review by Georgiou et al. [21].

OVERVIEW OF B CELL DEVELOPMENT AND ITS RELATIONSHIP TO ANTIBODY DIVERSITY

The B cell (*bursal* or bone marrow derived cell) is one of the many hallmarks of humoral immunity and its main function is to produce a diverse antibody repertoire in response to external stimuli and challenges [22]. Each B cell at any given moment can theoretically only produce a single unique antibody, thus targeting a unique cognate epitope. After a B cell is activated through cognate antigen binding, its differentiation toward the plasma cell state commences. The rate of antibody production continues to increase as an activated B cell differentiates into a plasma cell, where an enlarged endoplasmic reticulum is developed to accommodate higher antibody production. However, before a B cell commits to terminal differentiation to become a plasma cell, it is still transient in some ways, including with the potential ability to still modify the antibody it is producing [23]. Therefore, one can expect different stages of B cells would perform differently in terms of affinity and specificity [22]–[27]. Given the significance of B cell development, the following overview is provided and a schematic diagram summarizing the major B cell developmental stages is shown in Figure 4.

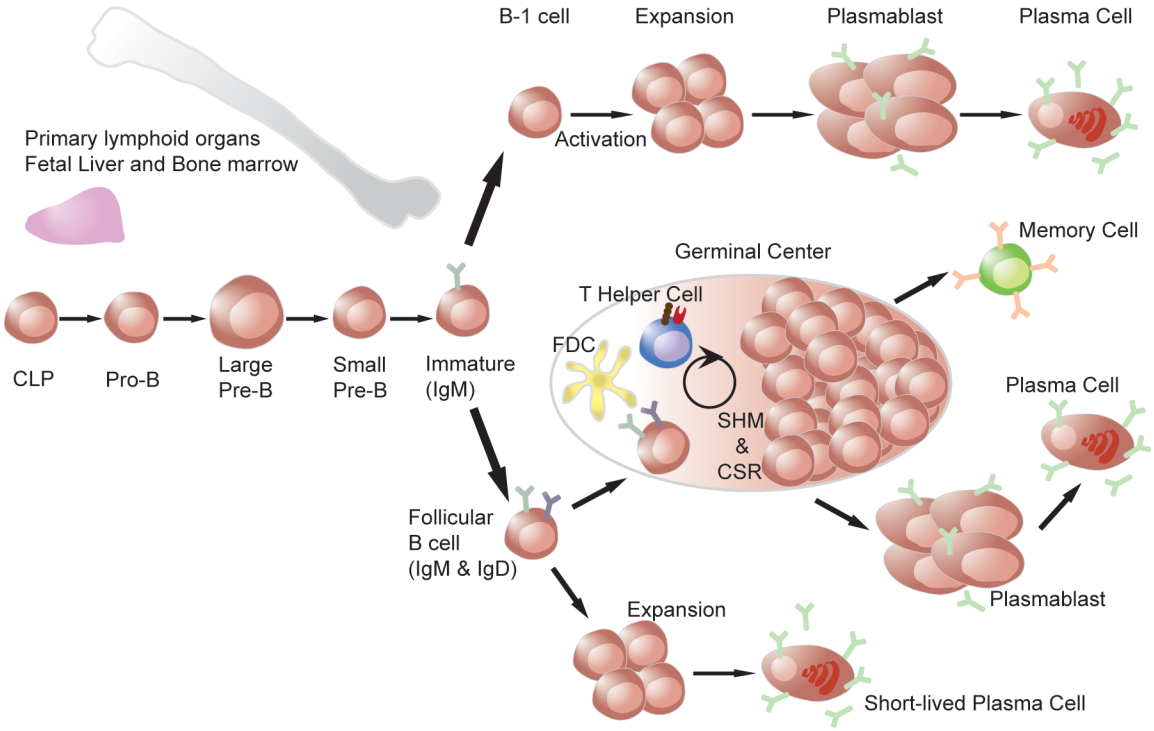

Figure 4: Brief schematic diagram of B cell development

Note: CLP = common lymphoid progenitor; FDC = follicular dendritic cell; SHM = somatic hyper mutation; CSR = class-switch recombination

B cells are derived from the common lymphoid progenitor (CLP) lineage and the initial stages of development occur in the fetal liver and subsequently in the bone marrow in neonatals and throughout adulthoods. During the pro-B cell stage, the pro-B cell is still undergoing imprecise variable domain gene rearrangement (which will be explained in more details in the next section) [28], [29]; therefore, the immunoglobulin is not expressed on the cell surface yet. Expression of an immunoglobulin chain is seen during the large pre-B cell stage where immunoglobulin-like structure or pre-B cell receptor (pre-BCR) can be found. The pre-BCR consists of a rearranged heavy chain with the μ isotype and surrogate light chain (SLC). The SLC is a generic light chain heterodimer

consisting of the VpreB and Lambda 5 components [30]. Given the error-prone nature of gene rearrangement, not all pro-B cells can transition to the pre-B stage by successfully forming a pre-BCR; hence, this is the first checkpoint for ensuring proper expression of immunoglobulin heavy chain. During the small pre-B stage, light chain gene rearrangements that result in successful expression can replace the SLC to form the surface immunoglobulin or B cell receptor (BCR), which signifies the immature B cell stage. During the immature stage, the BCR (surface IgM) is tested for auto-reactivity via a well-established biological mechanism [31]. If autoreactivity is detected then it is either corrected by a process called receptor editing or results in B cell apoptosis [31]–[33]. This checkpoint is important in preventing self-targeting B cells from being released into circulation. Upon passing this checkpoint, the immature B cell expresses surface IgD and is released from the primary lymphoid organ, whereupon the B cell is then classified as a mature B cell. Mature B cells further develop along two different lineages: B-1 or B-2 [34]. B-1 cells, after egressing from the bone marrow can mainly be found in pleural cavities and the peritoneum. They predominantly produce what are called “natural” antibodies (or innate antibodies), typically do not display SHM, are directed against T-independent (meaning not requiring T-cell costimulatory signals) antigens and their repertoire is restricted to the post fetal stage [35]. On the other hand, B-2 cells are mainly found in secondary lymphoid organs, such as the spleen and lymph nodes, where they produce adaptive antibodies mainly targeting T-dependent antigens. In terms of renewal, the B-1 cell is self-renewing whereas B-2 cells cannot proliferate (unless they are stimulated by cognate antigen binding, see below) and thus are continuously replenished via the influx of new cells from the bone marrow. B-2 cells are the major contributor to the diversity of antibody repertoire and will be discussed in more details hereafter.

As mentioned above, distinct B cell developmental stages ensure the generation and maintenance of antibody producing B cells. Mature B cells are estimated to encode for an astronomical diversity of unique antibodies with more than 10^{12} diverse clones [36]. A majority of the diversity can be found in the B-2 cell lineage where a large number of these cells reside near the follicles inside the secondary lymphoid organs (more specifically referred to as follicular B cells). Upon cognate recognition between the BCR and the antigens, these follicular B cells located within the follicles in the lymph nodes or spleen that have bound antigen become activated. Activation entails a rapid burst of cell expansion and differentiation to generate a nascent wave of plasmablasts (an intermediate stage of B cells to become plasma cells) and short-lived plasma cells that produce antibodies for immediate protection as depicted in the lower branch in Figure 4.

On the other hand, a portion of the activated B cells can refine their respective antibody's affinity and specificity in the germinal center as depicted in Figure 4. It is still unclear how this decision is made but it has been implicated that factors such as toll-like receptors and the type of antigen might affect the decision [37]–[39]. Germinal centers are compartmentalized sites in the secondary lymphoid organs where activated B cells undergo affinity maturation. A single germinal center can be divided into a light and a dark zone. The light zone is where follicular helper T cells (cognate T cells providing costimulatory signals), follicular dendritic cells (antigen presentation), and cognate B cells congregate. Follicular dendritic cells act as antigen presenting cells for the purpose of bringing cognate B cells in close proximity to follicular helper T cells. Due to a limited number of follicular helper T cells in the germinal center, their availabilities to provide B cell proliferative signals are restricted to only highly affinity matured cognate B cells. Cognate B cells receiving the proliferative signals can migrate and proliferate in the dark

zone where activation induced cytidine deaminase (AID) will introduce SHM to the antibody sequence. Independently, the AID reaction can also change the antibody's isotype through class-switch recombination (CSR) [25], [40] during this process. Through this cyclic selection process, only the B cells with affinity-refined antibody can be selected to remain in the germinal center for further refinement. These affinity-matured B cells either differentiate into memory B cells or into plasmablasts, although the underlying mechanisms that dictate this decision are unclear [41]–[48]. It is important to point out that long-lived memory B cells can also be generated independent of the germinal center [49]. Antibody producing B cells that have differentiated into plasmablasts have a short lifespan of a few days unless they can home to survival niches, mainly found in the bone marrow. Although still debatable, some scientists believed that plasmablasts need to localize into the bone marrow before terminally differentiating to plasma cells and to become long-lived plasma cells (LLPCs) [50]. LLPCs can survive up to at least several decades [51]. Immunological memory in the B cell branch is therefore maintained in the form of the memory B cell and LLPC populations.

It is well-studied that the bone marrow is the homing-destination for plasmablasts and that the bone marrow survival niches provide microenvironments for supporting long-term survival for non-dividing LLPCs [50]–[57]. Hence, the bone marrow LLPC pool is potentially an archival history of the antigenic specificities generated by the immune response; therefore, it plays a major role in immunological memory. Due to a limited number of survival niches [58], it is speculated that the diversity of the repertoire encoded by LLPCs is limited. Several studies reported that LLPCs can survive from as little as 60 days up until even the death of the host lasting decades [52], [56], [59]–[62].

OVERVIEW OF THE HUMAN IMMUNOGLOBULIN VARIABLE HEAVY/VARIABLE LIGHT LOCI RECOMBINATION DIVERSITY

The human immunoglobulin Variable Heavy chain gene locus is located on chromosome 14 (14q32.33) [63], while the Variable Light Kappa chain locus is located on chromosome 2 (2p11.2) [64], [65] and the Variable Light Lambda chain locus is located on chromosome 22 (22q11.2) [66]. It is necessary to join three gene segments (V-, D-, J-) for a heavy chain or two gene segments (V-, J-) for a light chain to form a productive full-length variable coding region for expression. Depending on which annotation database is referenced, the number of V-, D-, and J-segments may vary slightly. Nonetheless, according to the commonly used Immunogenetics Tools (IMGT) database [17], [67], [68], the Variable Heavy chain consists of 56 V-, 23 D-, and 6-J segments. The Variable Light Kappa chain has 38 V- and 5 J-segments while the Variable Light Lambda chain has 35 V- and 7-J segments. These segments are the fundamental genetic units for generating proficient diversity in the antibody repertoire. Somatic recombination joining of the V(D)J segments encode for the Variable Heavy chain and recombination joining of the VJ segments encode for the Variable Light chain. A schematic diagram depicting the sequential steps involved is shown in Figure 5.

Except for pseudo-gene segments, V-, (D- for heavy chain), J-, and constant gene segments are ordered as shown in Figure 5 on their corresponding chromosomal locus. With respect to the constant gene segments, in human the different isotypes are located neighboring the J-gene segments in the following order away from the centromere of the chromosome: Cμ, Cδ, Cγ3, Cγ1, ΨCε, Cα1, Cγ2, Cγ4, Cε, and Cα2 [63].

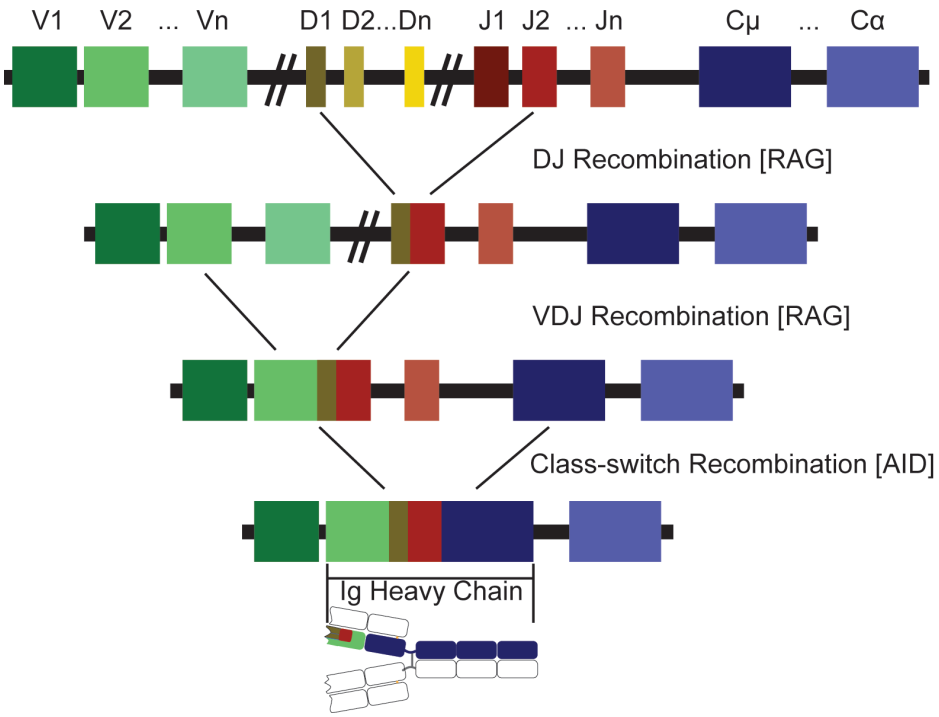

Figure 5: Schematic diagram of V(D)J rearrangement for Variable Heavy Chain

The joining of the V(D)J segments is mediated predominantly by the recombination-activating gene (RAG) complex (RAG-1 & RAG-2) but also by the Ku70:80 complex, the DNA-PK/Artemis complex, and by the terminal deoxynucleotidyl transferase (TdT). Briefly, the complexes can recognize the recombination signal sequences (RSSs) flanking the gene segments and bring them to a juxtaposition where imprecise excision removes sections of the chromosome to bring the to-be-joined segments into close proximity. This process leaves behind palindromic overhangs for each segment. TdT then introduces non-templated nucleotides (N-nucleotides) to the overhang. Afterwards, DNA repair enzymes correct and complete the complementary strand, leaving behind palindromic sequences (P-nucleotides). This process of N/P nucleotide addition signifies the completion of a gene segment-joining event. For heavy

chain rearrangement, the DJ joining precedes the V- and the constant segment joining. Since the light chain lacks the D-segment, the VJ joining precedes the constant segment joining.

If gene segment recombination were the only diversification strategy, the maximum theoretical number of Variable Heavy chain variants would only be around 7728 unique combinations. However, junctional diversity provides an additional layer of sequence diversity via non-templated (N/P) nucleotide addition resulting in incorporation of various numbers of nucleotides to the joining junctions. Thus, in theory, a diversity of more than 10^{12} of antibody genes is generated [36]. Together with combinatorial pairing between Variable Heavy and Variable Light chains, the theoretical diversity increases even more. However, some evidence using high throughput sequencing of V genes has suggested that the actual or true sequence diversity of the repertoire is much smaller and estimated at only 1-10 million clones [69]–[71]. The veracity of these estimates needs to be validated further. Nonetheless, a repertoire comprising less than 10^7 antibody sequences can be determined in its entirety by next generation sequencing (NGS) technology. As the operational costs of NGS continues to decline, it would be relatively practical to sequence at greater depth with comprehensive coverage.

NEXT GENERATION SEQUENCING TECHNOLOGY

For decades, the Sanger chain termination DNA sequencing technology has been the workhorse for DNA sequence analysis [72]. However, throughput is limited and large scale sequencing of DNA samples comprising billions of base pairs, as is the case for the immune repertoire is extremely labor intensive [73]. The greater throughput of NGS can

be attributed to innovation in the template preparation strategy, sequencing reaction methodologies, detection and informatics analysis. These innovative strategies enable the multiplexing of hundreds of thousands of DNA templates to be sequenced simultaneously [74].

Ever since the initial debut of NGS in 2005, the operational cost has continued to decline while the performance continues to improve rapidly. More importantly, substantial resources have been invested to improving accuracy, to provide longer read lengths, and for higher throughput [75]–[77]. Furthermore, with the huge potential market for personalized genome sequencing, there is additional incentive for the sequencer manufacturers to further reduce costs and to increase throughput making the products more enticing [74]. The emergence of many cutting-edge sequencing platforms, for example the Roche 454 pyrosequencing [78], [79], Ion Torrent PGM (Life Technologies), and Illumina MiSeq [80], have enabled high throughput analysis of several organisms' genomes, including the human in a reasonable amount of time [74]. In particular, the two platforms that are the most relevant in terms of the work presented in this dissertation are: the Roche 454 pyrosequencing and Illumina MiSeq. The Roche 454 pyrosequencing technology relies on emulsion PCR for template preparation [81] and pyrosequencing uses luciferase [78] as shown in Figure 6 top panel. Particularly, the Roche 454 pyrosequencing begins by clonal amplification in the oil-aqueous emulsion that isolates the PCR reaction of a primer conjugated bead with a single template molecule. The emulsion is disrupted after the clonal amplification resulting in each bead being covered with the monoclonal amplicon. These beads are then situated into plate with pico titer sized wells where the pyrosequencing steps will take place via the sequencing by synthesis method. Sequencing signals are generated when pre-defined

nucleotides are incorporated during the template polymerase reaction that will release pyrophosphate (PPi) stoichiometrically. The pyrophosphate together with adenosine 5' phosphosulfate (APS) can be quantitatively converted to ATP by sulfurylase. The stoichiometrically generated ATP can then power luciferase to convert luciferin to oxyluciferin emitting an amount of light proportional to the number of nucleotides incorporated. The light signal from each well is captured with a charge-coupled device (CCD) camera. With repeated cycles of washing and the introduction of pre-defined nucleotides, the sequence on each bead can be identified.

The Illumina MiSeq platform relies on solid-phase bridge PCR amplification and fluorescent-labeled reversible terminator to generate sequencing signals [80], [82] as shown in Figure 6 bottom panel. Briefly, common PCR primers are conjugated to a solid-phase glass slide (flowcell) where one cycle of PCR transfers the template onto the immobilized primers. Then, about 35 cycles of bridge PCR are used to generate clonal amplification in the form of an amplicon cluster. The amplicon cluster is generated in order to have sufficient DNA so that upon sequencing a strong signal can be readily detected. Specific cleavage of the strand in the opposite orientation ensures that each cluster only contains unidirectional strands. The sequences are read during strand synthesis where nucleotides extended from the sequencing primer are reported by the unique fluorescent-label on each type of reversible terminator nucleotide. In order to ensure single nucleotide extension, the nucleotides are originally blocked from further polymerization and only after nucleotide signals are captured by a total internal reflection fluorescence (TIRF) detector are the incorporated nucleotides unblocked to allow further extension. After repeated rounds of nucleotide incorporation, signal detection, and nucleotide unblocking, the representative sequence from each cluster is revealed.

The Roche 454 pyrosequencing and Illumina sequencing platforms each have their own respective pros and cons. For example, the Roche 454 pyrosequencing platform offers single read-throughs of ~750 bps in length but has at maximum sequencing depth of only 1 million reads per run. On the other hand, Illumina MiSeq can generate up to about 25 million reads per run but the read length is restricted to only 300 bps on each pair end currently. In order to ascertain longer read lengths with the MiSeq system, pair-end reads are required with sufficient overlap for *in silico* assembly algorithms to "stitch" together each end at the expense of half the sequencing depth. The turnaround time for Illumina MiSeq is also longer where a typical run requires about 40 hours whereas the turnaround time for the Roche 454 pyrosequencing is only 24 hours. In terms of sequencing errors, the Roche 454 pyrosequencing platform is prone to generate insertion/deletion errors in homopolymer regions whereas the Illumina MiSeq platform tends to generate substitution errors for bases after guanine. Regardless, both platforms are currently the best possible options for surveying antibody repertoires at a reasonably high depth and at relatively low cost.

The applications of NGS ranges from genome assembly to transcriptome analysis to targeted amplicon sequencing, etc. In particular, NGS can simultaneously read antibody sequences in the range of millions resulting in a comprehensive measure of the repertoire. This ability has been capitalized to find antigen specific antibodies [83]–[85] and to identify repertoire signatures in healthy individuals or in individuals with immune malignancies [69], [70]. Moreover, it could be used to pin point self-reactive antibodies in rheumatoid arthritis patients [86]. Hence, NGS is utilized throughout the work

described in this dissertation to survey the antibody repertoire in ways that could lead to implications in antibody discovery or identification of other repertoire signatures.

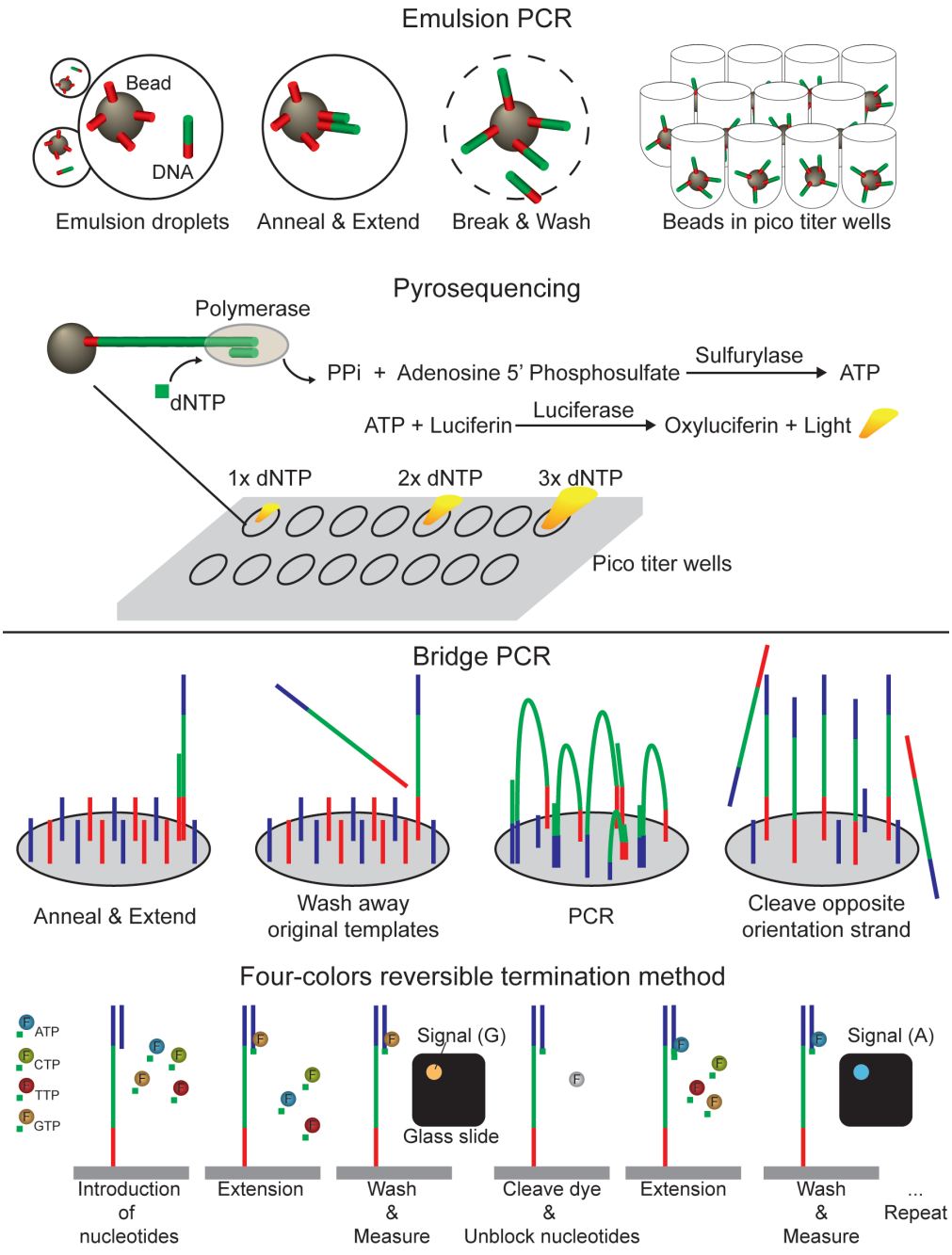

Figure 6: Schematic diagram for 454 pyrosequencing and Illumina MiSeq sequencing

BIOINFORMATICS AND EXPERIMENTAL TOOLS FOR ANTIBODY REPERTOIRE STUDIES

As mentioned previously, the ability for NGS to measure the antibody repertoire can lead to medical implications for therapeutics or diagnostics. In order to comprehensively determine the sequences that comprise the antibody repertoire, it is necessary to use a multi-layer approach by combining NGS, bioinformatics tools, immunological assays, and recombinant antibody technology into an integrative pipeline. Bioinformatics tools and analysis remain a critical bottleneck to the extraction of important technical and biomedical information from NGS sequencing datasets, especially when the influx of data is far greater than any manual processes can handle. Moreover, it is essential to maintain a low computational skill barrier for widespread adoption of the approach. Therefore, many bioinformatics-centric groups have published various packages and analysis tools to lower the level of computational skill required for the analyzing of NGS data [87]–[96].

The bioinformatics tools related to antibody repertoire analysis that were made publicly available can be categorized into four groups: pre-processing tools, germline annotation tools, clustering tools, and visualization tools. The respective web locations for the tools are summarized in Table 1.

Pre-processing tools	
FASTX toolkits	http://hannonlab.cshl.edu/fastx_toolkit/
Ig-HTS tools	http://immsilico2.lnx.biu.ac.il/Software.html
Flash	http://ccb.jhu.edu/software/FLASH/
PEAR	http://sco.h-its.org/exelixis/web/software/pear/
Germline annotation tools	
IMGT/V-QUEST	http://www.imgt.org/HighV-QUEST/index.action
IgAT	http://www.uni-marburg.de/fb20/kinderklinik/forschung/neonat/igat
IgBlast	http://www.ncbi.nlm.nih.gov/igblast/
SoDA2	http://dulci.biostat.duke.edu/computationalimmunology/antigen/soda2.html
iHMMune-align	http://www.emi.unsw.edu.au/~ihmmune/index.php
VDJsolver	http://www.cbs.dtu.dk/services/VDJsolver/
Ab-orgin	http://mpsq.biosino.org/ab-origin/supplementary.html
Antibody mining toolbox	http://sourceforge.net/projects/abmining/
Clustering tools	
Usearch	http://www.drive5.com/usearch/
CD-HIT	http://weizhong-lab.ucsd.edu/cd-hit/
ClonalRelate	http://www.cse.unsw.edu.au/~ihmmune/ClonalRelate/
Visualization tools	
IgTree	http://immsilico2.lnx.biu.ac.il/Software.html
Circos	http://circos.ca/

Table 1: Summary of bioinformatics tools and their web locations

Sequence quality could vary from read-to-read and it is common practice for researchers to first prune irrelevant and low-quality sequences. Pre-processing tools designed for this task are usually made available in a programming language such as PERL. Two commonly used tools in this area are FASTX Toolkit from the Hannon Laboratory and the Ig-HTS tool [97] by the Mehr Laboratory. These tools can automate multiplex tag trimming and the filtering of short/low-quality reads. In addition to sequence pruning, sequence merging is also necessary for analyzing antibody sequences identified through the Illumina MiSeq platform. Since the MiSeq run currently only supports up to 300 bps of read length from each end of a template molecule, recovery of the full-length variable region requires merging of the pair-end reads. Tools such as FLASH [98] and PEAR [99] are exceptionally efficient in the merging of paired-end reads. Briefly, on the one hand, FLASH's algorithm is meant to minimize the overlap length to mismatch ratio and it requires mean DNA fragment size and standard deviation of fragment population as input parameters to generate merged reads that are almost identical in size. On the other hand, PEAR utilizes an overlap assembly score based on a tuned scoring matrix that heavily penalizes mismatches and rewards matches to generate merged reads. In our experience, both tools perform equally well in most cases but on rare occasions, PEAR can outperform FLASH in recovering more merged reads.

Among the different tools, the germline annotation tools that can assign the V-, D-, J- gene segments are of great importance because such information can be used to intrinsically infer the ancestry of the B cells and identify certain gene usages that have been found to be associated with certain diseases [100]–[103]. There are various tools that automate the annotation process but IMGT/V-QUEST by far appears to be the one that has been the most widely adopted [89]. IMGT/V-QUEST offers a web-based

platform using a proprietary annotation scheme to report the annotations along with detailed periphery information such as somatic mutation levels and isoelectric points organized across multiple tab-delimited files. These files can then be subsequently processed and transformed by downstream programs such as the Immunoglobulin Analysis Tool (IgAT) [104] to extract repertoire information such as CDR3 spectra counts and N/P addition analysis. The downside of IMGT/V-QUEST is that it imposes a limit of 500,000 sequences per analysis making analysis in the tens of millions quite inconvenient. The next commonly used annotation tool is IgBlast [92] that is implemented using the National Center for Biotechnology Information (NCBI) C++ Blast toolkit and reports results in the GenBank alignment format. The information reported by IgBlast is not as easily extracted and, more importantly, it does not report the CDR3 region which is often times a critical piece of information. SoDA2 [91] and iHMMune-align [88] are both based on hidden Markov models for annotation and they are deployed under a JAVA standalone package. Their processing capabilities seemed to be limited to the range of 10^6 sequences. VDJsolver [105] is a combination of JointHMM (hidden Markov model) and JointML (maximum-likelihood) based methods to obtain the best fit to a typical antibody model. Ab-origin [90] optimized a scoring scheme based on a Monte Carlo simulation that minimizes effects from large length variations in the V(D)J joint. Lastly, the Antibody mining toolbox [93] uses regular expression pattern finding of conserved motifs to identify the CDRs within the variable gene sequence. As observed above, the underlying concept for the annotation tools is the ability to map the antibody sequence with a degree of flexibility to the germline database. Results generated by each tool can be affected by the alignment approach. Hence, until there is a standard set from which to benchmark the tools, there is still no clear consensus as to which tool is more

accurate. However, researchers tend to welcome the tool that provides the most periphery information; therefore, IMGT/V-QUEST is still popular among immunologists.

In addition to annotation tools, tools that can cluster sequences into clonotypic units are also critical in determining antibody ancestry relationship. The ones that are commonly used are Usearch [94], CD-HIT [95], and ClonalRelate [96]. Last but not least, visualization tool like IgTree [106] that constructs phylogenetic trees can illustrate the maturation path of an antibody sequence while a Circos plot [107] can illustrate shared sequences in different sample populations.

SUMMARY

The dominance of mAbs as effective therapeutics is expected to remain in the coming decades. And, the NGS technology has presented us with unprecedented high throughput ability to mine the antibody repertoire in different disease states or during vaccination regime that can reveal great amount of information in advancing immunotherapy. By combining bioinformatics processing and analysis tools with NGS technology, we demonstrate in this dissertation the utility of such integrative approach in antibody discovery, immune traits identification in healthy adults, determination of new germline and gene conversion rates in rabbit, and finally the characterization of circle sequencing in expanding the sequencing quality.

Chapter 2: Isolation of Monoclonal Antibodies without Screening by Mining the Variable Gene Repertoire of Plasma Cells

Specifically for the work described in this chapter, K.H. Hoi performed the immunization and bone marrow sample collection together with S.T. Reddy. K.H. Hoi has also performed bacterial cloning and expression of the antibodies (scFvs and full-length), the reagents used in the experimental assays (Western blots, ELISAs and immune-precipitation). In terms of bioinformatics analysis, K.H. Hoi has improved the initial script X. Ge created to increase performance and functionality for reporting additional information regarding the antibody sequencing data.

This chapter is reproduced with minor modifications from its initial publication:
S. T. Reddy, X. Ge, A. E. Miklos, R. A. Hughes, S. H. Kang, K. H. Hoi, C. Chrysostomou, S. P. Hunicke-Smith, B. L. Iverson, H. O. Tucker, A. D. Ellington, and G. Georgiou, "Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells," *Nat Biotech*, vol. 28, no. 9, pp. 965–969, 2010.

Manuscript author contributions:
S.T. Reddy and G. Georgiou developed the methodology, designed the experiments, analyzed the data, and wrote the manuscript; X. Ge, C. Chrsostomou, and K.H. Hoi carried out the bioinformatics analysis; S.T. Reddy, A.E. Miklos, R.A. Hughes, S.H. Kang, and K.H. Hoi performed the experiments; S.P. Hunicke-Smith performed 454 sequencing; B.L. Iverson, H.O. Tucker, and A.D. Ellington analyzed the data.

Acknowledgements:
The authors would like to acknowledge Chyya Das for her technical assistance on mammalian cell culture and transfection and Michelle Byrom for her assistance with mouse experiments. This work was funded by grants from the Clayton Foundation, and the Cockrell Chair in Engineering to G. Georgiou and by a fellowship from Natural Sciences and Engineering Research Council of Canada to X. Ge.

INTRODUCTION

The ability of the mammalian humoral immune response to generate a vastly diverse antibody repertoire as a response to stimuli (i.e., immunization) has been substantially exploited for biotechnology applications, specifically, monoclonal antibody

(mAb) isolation [108]. Since the development of the hybridoma technology by Kohler and Milstein 35 years ago [109], a variety of methods for the generation of mAbs have been developed. Such methods include B cell immortalization by genetic reprogramming via Epstein-Barr Virus (EBV) [110] or retrovirus-mediated gene transfer [111], cloning of V genes by single cell PCR [112], [113], and methods for *in vitro* discovery via the display and screening of recombinant antibody libraries [114]–[121]. Both *in vitro* and *in vivo* methods for antibody discovery are critically dependent on high-throughput screening to determine antigen specificity. Recently, B cell analysis has been expedited by microengraving techniques that utilize soft lithography for the high-throughput identification of antigen-specific B cells [119], [120], however, this is at the cost of considerable technical complexity due to the need for antibody variable gene amplification and cell expansion. Similarly, the success of *in vitro* antibody discovery techniques is dependent on screening parameters including the nature of the display platform, antigen concentration, binding avidity during enrichment, multiple rounds of screening (e.g, panning or sorting), and importantly, on the design and diversity of synthetic antibody libraries [108], [121], [122].

In an effort to streamline the process, we developed a simple and rapid method for antibody isolation without the need for any laborious screening steps. We exploited high-throughput DNA sequencing to analyze the variable light chain (V_L) and variable heavy chain (V_H) antibody gene repertoires derived from the mRNA transcripts of fully differentiated mature B cells, also known as antibody secreting plasma cells, found within the bone marrow of immunized mice. Following bioinformatics analysis, several abundant and unique antibody V_L and V_H gene sequences could be identified within the repertoire of each immunized mouse. By utilizing the automated liquid handling robots,

synthetic antibody genes were rapidly generated by oligonucleotide and PCR assembly. Antibodies were recombinantly expressed in bacterial and mammalian systems as single-chain variable fragments (scFv) in the bacterial system and full-length IgG in the mammalian system (Figure 7). The pairing of V_L and V_H genes was accurately predicted by representation within the repertoire and it was confirmed that abundant and unique sequences corresponded to antigen-specific antibodies, thus providing a method for rapid and direct isolation of mAbs without screening.

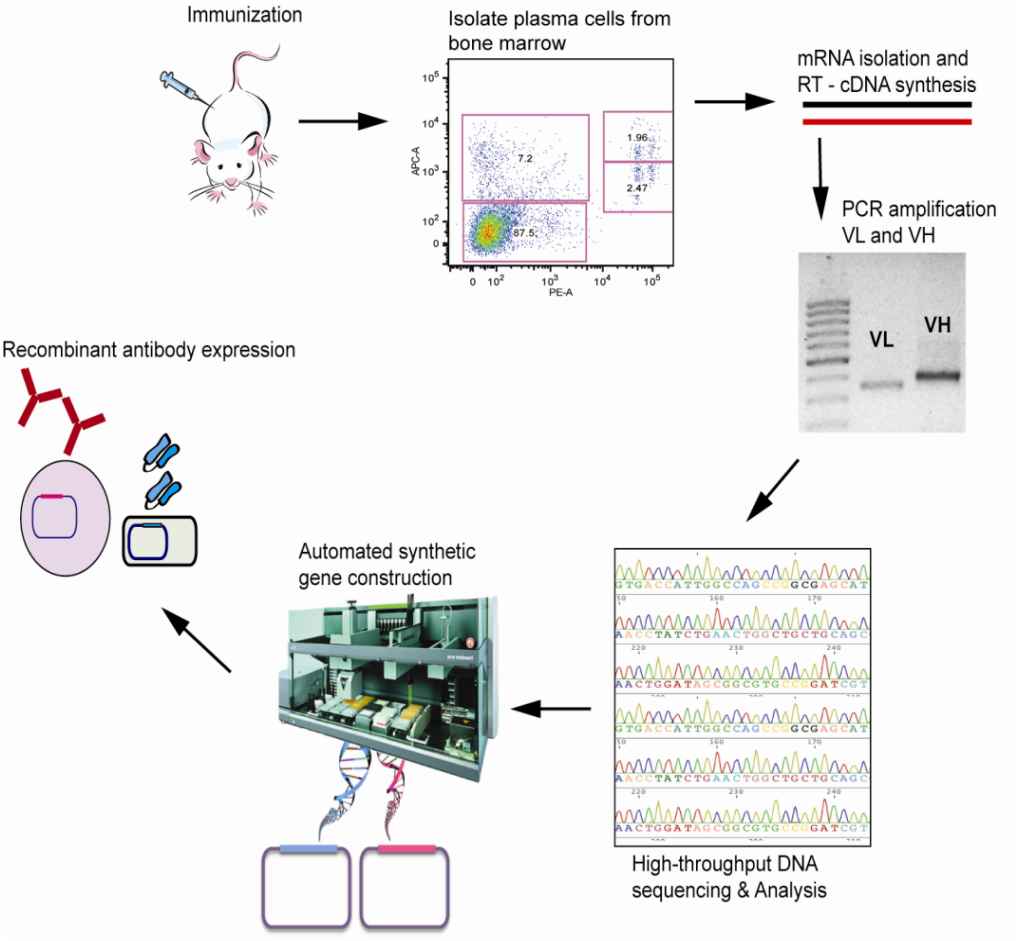

Figure 7: Schematic for isolation of monoclonal antibodies without screening by mining the antibody variable (V) gene repertoires of bone marrow plasma cells

Moreover, B cell maturation culminates in the terminal, non-proliferative stage of B cell development - the formation of plasma cells that serve as immunoglobulin production factories. Plasma cells represent less than 1% of all lymphoid cells and yet are responsible for the overwhelming majority of antibodies in circulation [123], [124]. The bone marrow constitutes the major compartment where plasma cells take residency and produce antibodies for prolonged periods of time. In mice, a stable and highly-enriched antigen-specific BM-PC population of ~10^5 cells (10-20% of all BM-PCs) appears 7 days following secondary immunization and persists for prolonged periods [60]. In contrast, the splenic plasma cell population is highly transient, as it peaks at day 7 and rapidly declines to <10^4 cells by day 11. Importantly, BM-PCs are responsible for the synthesis of the most abundant circulating antibodies which in turn are likely to play a dominant role in pathogen neutralization and other protective humoral immune responses [123]. Hence, in order to examine the dynamics of the antibody V gene repertoires in BM-PCs, especially during the early phase post challenge (i.e., to mimic situations where mice exhibit weak immune responses), pairs of mice were immunized with chicken egg ovalbumin (OVA), human complement serine protease (C1s), human B cell regulator of IgH transcription (Bright), or adjuvant only as control to explore the differences in dynamics under different challenging antigens.

MATERIALS AND METHODS

Immunization

Purified human complement protein C1s (CalBiochem), purified chicken egg ovalbumin (OVA, Sigma), or recombinant bacterially-expressed human B cell regulator

of IgH transcription (Bright) were resuspended in sterile-filtered phosphate buffered saline (PBS) at 1.0 mg/ml. On the day of primary immunization, 25 ul of antigen solution was thoroughly mixed with 25 ul of Complete Freund's Adjuvant (CFA, Pierce Biotechnology) and 50 ul of sterile PBS and stored on ice. Female Balb/c mice (Charles Rivers Laboratories) 6-8 weeks old were housed in conventional barrier space and were maintained on a normal chow diet. Prior to injections, mice were bled from the tail vein and approximately 25 ul of blood was collected and stored at -20°C for later analysis. Day 1 was designated as the day primary immunizations were performed. 100 ul of the antigen-CFA mixture per mouse was injected with a 26-gauge needle subcutaneously into the backpad. Mice were monitored daily by animal housing staff and cages were changed twice per week. For secondary immunization, 25 ul of antigen solution was thoroughly mixed with 25 ul of Incomplete Freund's Adjuvant (IFA, Pierce Biotechnology) and 50 ul of sterile PBS and stored on ice. On day 21 mice were given the secondary immunization intraperitoneally at 100 ul of antigen-IFA mixture per mouse. On day 26 mice were sacrificed by CO_2 asphyxiation and blood, femurs and tibia were collected. All experiments were conducted following the guidelines provided by the university's Institutional Animal Care and Use Committee (protocol number AUP-2009-00016).

Isolation of bone marrow plasma cells

Muscle and fat tissue was removed from the harvested tibias and femurs. The ends of both tibia and femurs were clipped with surgical scissors and bone marrow was flushed out with a 26-gauge insulin syringe (Becton Dickinson, BD). Bone marrow tissue was collected in sterile-filtered Buffer#1 (PBS with 0.1% bovine serum albumin (BSA)/2 mM ethylenediaminetetracetic acid (EDTA)). Bone marrow cells were collected by

filtration through a 70-um cell strainer (BD) with mechanical disruption and washed with 20 ml of PBS and collected in a 50 ml tube (Falcon, BD). Bone marrow cells were then centrifuged at 1200 RPM for 10 min at 4°C. Supernatant was decanted and the cell pellet was resuspended with 3.0 ml of red blood cell lysis buffer (eBioscience) and shaken gently at 25°C for 5 minutes. Cell suspension was then diluted with 20 ml of PBS and centrifuged at 1200 RPM for 10 minutes at 4°C. Supernatant was decanted and cell pellet was resuspended in 1.0 ml of Buffer#1.

Each isolated bone marrow cell suspension was incubated with 2.5 ug and 1.5 ug of biotinylated rat anti-mouse CD45R(B220) and biotinylated rat anti-mouse CD49b (eBioscience), respectively. Cell suspension was rotated at 4°C for 20 minutes. Cell suspensions were then centrifuged at 2,000 RPM for 6 minutes at 4°C, supernatant was removed and the cell pellet was resuspended in 1.5 ml of Buffer#1. Streptavidin conjugated M280 magnetic beads (Invitrogen) were washed and resuspended according to manufacturer's protocol. 50 ul of magnetic beads were added to each cell suspension and the mixture was rotated at 4°C for 20 min. Cell suspensions were then placed on Dynabead magnet (Invitrogen) and supernatants (negative fraction, cells unconjugated to beads) were collected and cells bound to beads were discarded.

Pre-washed streptavidin M280 magnetic beads were incubated for 30 min at 4°C with biotinylated rat anti-mouse CD138 (BD Pharmingen) with 0.75 ug antibody per 25 ul of magnetic beads. Beads were then washed according to manufacturer's protocol and resuspended in Buffer#1. The negative cell fraction (depleted of $CD45R^+$ and $CD49b^+$ cells) collected as above was incubated with 50 ul of CD138 conjugated magnetic beads

and the suspension rotated at 4°C for 30 min. Beads with $CD138^+$ bound cells were isolated by the magnet, washed 3 times with Buffer#1, the negative (CD138-) cells unbound to beads were discarded (or saved only for analysis). The positive $CD138^+$ bead-bound cells were collected and stored at 4°C until further processed.

Preparation of variable light chain V_L and variable heavy chain V_H genes

$CD45R^-CD138^+$ BM-PCs isolated as described herein were centrifuged at 2,000 RPM at 4°C for 5 min. Cells were then lysed with TRI reagent and total RNA was isolated according to the manufacturer's protocol in the Ribopure RNA isolation kit (Ambion). mRNA was isolated from total RNA with oligodT resin and the Poly(A) purist kit (Ambion) according to the manufacturer's protocol. mRNA concentration was measured with an ND-1000 spectrophotometer (Nanodrop).

The isolated mRNA was used for first strand cDNA synthesis by reverse transcription with the Maloney murine leukemia virus reverse transcriptase (MMLV-RT, Ambion). cDNA synthesis was performed by RT-PCR using 50 ng of mRNA template and oligo(dT) primers according to manufacturer protocol of Retroscript kit (Ambion). Following cDNA construction, PCR amplification was performed to amplify the V_L and V_H genes using 2 ul of unpurified cDNA product and standard V_L and V_H primer mixtures [125]. PCR products genes were gel purified and submitted to Genomic Sequencing and Analysis Center at the University of Texas Austin for 454 DNA sequencing.

High-throughput sequencing of V_H and V_L repertoires

V gene repertoires isolated from BM-PC of eight mice were sequenced using high-throughput 454 FLX sequencing (University of Texas, Austin, TX; SeqWright, Houston, TX). In total, 415,018 sequences were generated, and 454 data quality control filtered and grouped >97% of the sequences into datasets for each mouse according to their Multiplex Identifiers (MID) usages.

Bioinformatics analysis of V gene repertoires

CDR3 sequences were identified based on homology to conserved flanking sequence motifs found upstream and downstream to CDR3s. Searching motifs for CDRH3 and CDRL3 were determined based on amino acids which occur with an average frequency of 99% at specific positions in V genes, based on the Kabat database [16]. V_H sequences were identified by searching for DXXX(Y/F)(Y/F)C (Kabat # 86-92) and WGXG(T/S) (Kabat # 103-107) motifs at the N- and C- termini of CDRH3 respectively. Analogously, V_L genes were found by searching for degenerate codon sequences encoding DXXXY[F/Y]C (Kabat # 82-88) and FGXGT (Kabat # 98-102). This approach correctly identifies over 94% of V_H and 92% of full-length V_L sequences in the Kabat database. Any sequences or reverse complements containing these motifs encoding in-frame CDR3 sequences were extracted as putative V_H or V_L genes. For each sample, CDR3 sequence abundance and frequencies were calculated. V genes containing high frequency CDR3s were used for BLAST searches of the sequence database. Full-length V gene sequences were accepted if they covered all 3 CDRs. Subsequently, pairwise homologies among full-length sequences were determined using the multiple sequence alignment tools (Geneious Software). Analyses were performed using Perl scripts in a

Unix environment and converted into a graphical user interface using Matlab 7.1 for user-friendly environment.

Construction of synthetic antibody genes

The coding sequences for the selected V_L and V_H genes were designed using the GeneFab software component of our in-house protein fabrication automation (PFA) platform [126]. After reverse translation of the primary amino acid sequences for each V_L and V_H using an *E. coli* class II codon table, the coding sequences for each V_L and V_H were paired based upon the rank abundances and relative frequency from the sequencing data (most abundant V_L with the most abundant V_H so and so forth). The antibody V region sequences were built in scFv format with a poly-glycine-serine linker $(GGGGS)_4$ between the V_L and V_H sequences. The rank ordered scFv sequences were then aligned using the common $(GGGGS)_4$ linker sequence and a universal randomly generated stuffer sequence was applied to the ends of the scFv sequences to ensure that all of the synthesized constructs were the same length (808bp). This design format reduced the number of oligonucleotides needed for gene synthesis as oligonucleotides with identical sequences between the different scFv constructs could be reused. SfiI restriction endonuclease sites were added flanking each gene sequence to facilitate cloning of the synthetic gene constructs into compatible pMoPac vectors [127].

The scFv genes were synthesized from overlapping oligonucleotides using a modified thermodynamically balanced inside-out nucleation PCR protocol [128]. The 80-mer oligonucleotides necessary for the construction of the various scFv genes were designed using the GeneFab software with a minimal overlap of 30 nucleotides between

oligonucleotide fragments. The oligonucleotides were synthesized using standard phosphoramidite chemistry at a 50nmol scale using a Mermade 192 oligonucleotide synthesizer (Bioautomation; Plano, TX) using synthesis reagents from EMD Chemical (Gibbstown, NJ) and phosphoramidites from Glen Research (Sterling, VA). All of the oligonucleotide liquid handling operations necessary for assembling the various genes were done on a Tecan Evo 200 workstation (Tecan; Mannedorf, Switzerland) with reagent management and instrument control done through the FabMgr software component of the PFA platform[2]. The gene assembly PCRs were performed using KOD-Hotstart polymerase using buffers and reagents supplied with the enzyme (Novagen; San Diego, CA). To facilitate cloning of the V_L and V_H genes separately into vectors for IgG expression, the genes for the various V_L and V_H pairs were either built as gene fusions similar to the scFvs except without the $(GGGGS)_4$ linker or as separate genes. These constructs contained sites for the restriction enzymes BssHII and BsiWI flanking the V_L gene and the BssHII and NheI sites flanking the V_H gene.

Antibody expression and antigen binding analysis

Antibody fragments were expressed as scFv fusions to the human light chain constant region Cκ (scAbs) [127], which possessed a by a C-terminal a polyhistidine (polyHis) tag. Cloning was accomplished by SfiI digestion of antibody genes and ligation into the expression vector pMoPac16 followed by electroporation transformation into *E.coli Jude 1* cells, which were then plated on Luria Broth (LB, Miller) agar plates supplemented with 100 ug/ml ampicillin. Single colonies were used to inoculate cultures in microtiter 96 well plates with 200 ul/well of Terrific Broth (TB, Miller) supplemented with 2% glucose and 100 ug/ml ampicillin; plates were shaken for 16 h at 30°C. 10 ul of

each well was used to inoculate 200 ul/well of fresh 96 well plates containing TB media supplemented with 100 ug/ml ampicillin and 1 mM of Isopropyl-β-D-thiogalactopyranoside (IPTG, Calbiochem).

Following a 4h induction, plates were centrifuged at 4,000 RPM for 10 min at 4°C, the supernatent was decanted and cell pellets were resuspended in 20% BugBuster HT (Novagen) in PBS at 150 ul/well. Plates were then shaken at 25°C for 30 min, and then centrifuged at 4,000 RPM for 15 min at 4°C. 50 ul/well of cell lysates were then added to an ELISA 96 well plate that was pre-coated with antigen (e.g., OVA, C1s, Bright) at 2 ug/ml in PBS and pre-blocked with 0.5% BSA or 1% Gelatin. A standard indirect ELISA protocol was followed with the detection anti-polyHis antibody (Sigma) conjugated to horseradish peroxidase (HRP) and developed with TMB substrate (Dako) for 15-45 minutes and stopped with 2N H_2SO_4. The absorbance was measured at 450 nm with a 96 well spectrophotometer (BioTek). Positive wells were identified when the absorbance value was at least 3-fold above background binding to BSA.

For IgG expression, synthetic V_L and V_H genes were digeseted with *BssHII*/*BsiWI* and *BssHII*/*NheI* respectively and then ligated into the vectors pMAZ-IgL and pMAZ-IgH, respectively [129]. pMAZ-IgL carries the constant human kappa light chain antibody region and pMAZ-IgH carries the constant human heavy chain antibody region of IgG1. Vectors were transformed into *E.coli Jude 1* cells and plated on Luria Broth (LB, Miller) agar plates supplemented with 100 ug/ml ampicillin. Single colonies were selected and verified for correct V gene sequence. *E.coli* cells carrying pMAZ-IgL and pMAZ-IgH vectors were then grown in 2 ml TB supplemented with 100 ug/ml ampicillin isolated and DNA was purified. 20 ug each of purified pMAZ-IgL and pMAZ-IgH were

used for co-transfection and transient expression from HEK293F cells following the Freestyle MAX expression system (Invitrogen). HEK293F cells were grown for 96h following transfection and media was harvested and IgG was purified by a protein-A agarose chromatography column.

Surface Plasmon Resonance (Biacore)

C1s was covalently immobilized on a CM5 chip (GE healthcare, NJ) at a level of approximately 200 response units via standard amine coupling chemistry as described in manufacturer's protocol. BSA was similarly coupled for baseline correction. All kinetic analyses were performed at 25℃ in HBS-EP (10mM HEPES, 150mM NaCl, 50μM EDTA, 0.005% P-20, pH 7.4) on a BIAcore 3000 (GE healthcare, NJ). Antibodies were injected over immobilized antigen at a flow rate of 50μl/min or 100μl/min and the chip was regenerated with a single 10s injection of 20mM NaOH. Each sensogram was run in duplicate. Kinetic and equilibrium constants were determined by global fitting to a bivalent model using BIAevaluation software (GE healthcare, NJ).

RESULTS

Unlike recent high-throughput sequencing analyses that explored V gene repertoire diversity in zebrafish, humans, or synthetic libraries [69], [70], [130], [131], our goals focused on: determining the relative V gene transcript abundance in the BM-PC repertoires of immunized mice and that highly abundance transcripts to be likely antigen specific. These tasks do not require exhaustive coverage of the V gene repertoire; we have found that obtaining >5k V gene sequences per BM-PC sample is sufficient to provide the information needed for antibody discovery, minimizing DNA sequencing

costs. 454 reads were first processed by multiple sequence filters, and then subjected to a simple and rapid bioinformatics analysis that relied on homologies to conserved framework regions within V genes to identify the most common complementarity determining region 3 (CDR3) sequences. This approach correctly identified ~94% of V_H and ~92% of V_L sequences in the Kabat database [16]. Out of a total of 415,018 reads, 23.2% contained CDRH3 and 26.6% contained CDRL3 sequences (Table 2), representing 6,681-16,743 CDRH3 and 7,112-21,241 CDRL3 sequences read per mouse, respectively.

Sample	454 GS-FLX Sequencing Size	Sequences Containing CDR-H3 motif			Sequences Containing CDR-L3 motif		
		Total Number	Number of Unique CDR-H3	Number of CDR-H3 as Single Copy	Total Number	Number of Unique CDR-L3	Number of CDR-L3 as Single Copy
Adjuvant-1	32066	6681	2706	1811	7112	1638	1053
Adjuvant-2	86720	16743	4640	2890	21241	3136	1888
Ova-1	63872	15350	4789	3010	13355	2251	1355
Ova-2	72257	15751	3821	2401	17200	2786	1700
C1s-1	43753	11595	2440	1443	13972	1706	1045
C1s-2	39961	9071	1799	999	14664	1477	847
Bright-1	36599	9453	2025	1178	12209	1383	632
Bright-2	39790	11769	2530	1210	10441	1422	578
Total	**415018**	**96413**	**24750**	**14942**	**110194**	**15799**	**9098**
Unique Sequences Across All Samples			**21271**			**8690**	

Table 2: Reads summary for 454 DNA sequences containing CDR3

V gene sequences containing a particular CDR3 were accepted as full-length if they covered all 3 CDRs. Pairwise identities and frequencies were calculated by multiple sequence alignments followed by germline analysis. Concurrently, a graphical user interface application was developed to enhance data analysis and visualization of the results. Analysis of the BM-PC repertoires led to several interesting observations. First, in all immunized mice, including those receiving the same antigen, >92% of the CDRH3

sequences were unique to an individual mouse. The CDRL3 repertoires were less diverse, and in some instances, BM-PCs from mice immunized with different antigens expressed high levels of the same CDRL3 (data not shown). A lower degree of V_L diversity, especially in early responses (as was the case here) is logical, since CDRL3 is derived from a single V-J recombination as opposed to two recombination events (V-D-J) for CDRH3. Second, and most importantly ~10-20% of the total repertoire of all immunized mice were on average comprised of only 4 CDRH3 sequences (data not shown). For example, in the two mice receiving C1s, the frequencies of the most abundant CDRH3s were 7.93% and 10.99% of the total repertoire. Third, as expected for early responses, the most highly abundant CDR3s were assembled from a diverse array of germline V gene segments, with average somatic mutation rate of only 2 and 5 amino acid substitutions for V_L and V_H, respectively. Not surprisingly, certain germline V gene families were represented preferentially in mice responding to particular antigens. For example in mice immunized with C1s, between 15-30% of the entire V_H gene repertoire utilized IGHV1 family whereas the adjuvant only immunized mice were dominated by IGHV5 or IGHV6 families.

In most instances the V genes encoding a highly abundant CDR3 were dominated by one sequence with the second most abundant V gene sequence (somatic variant) being present at >10-fold lower level and differing from the dominant sequence by 1-2 amino acids (Table 3), where the first number after the antigen represents the animal number and the second number after the period represents the rank of the sequence.

Antigen	CDR3	CDR3 Freq (%)	1st V_H Freq (%)[a]	2nd V_H Freq (%)[a]	V_H Homology[b]
OVA-1.1	GSSYYAMDY	7.11	60.0	1.7	96.1
OVA-1.2	DYYGSSYWYFDV	1.10	47.1	5.8	89.9
OVA-1.3	DNWDWYFDV	0.57	49.0	4.0	95.0
OVA-1.4	LLWLYAMDY	0.54	54.7	4.7	97.3
OVA-2.1	RTTVSRDWYFDV	7.61	15.3	5.6	92.3
OVA-2.2	YYYGSSAMDY	3.23	26.0	10.8	96.0
OVA-2.3	DGWYYFDY	2.22	22.7	4.1	89.1
OVA-2.4	EDDYDLFAY	2.10	9.4	8.7	94.9
C1s-1.1	GNYYYAMDY	7.93	68.8	1.1	97.9
C1s-1.2	DDGYWYFDV	5.14	60.9	5.3	90.0
C1s-1.3	YYYGSSAMDY	4.37	58.5	3.7	94.5
C1s-1.4	DMISYWYFDV	2.64	70.9	1.1	90.0
C1s-2.1	SDRYDGYFDY	10.99	11.1	9.4	95.7
C1s-2.2	SDRFDGYFDY	9.93	12.5	4.2	94.7
C1s-2.3	WLLLAY	3.30	26.3	7.7	88.8
C1s-2.4	YGNYFDY	2.47	72.1	1.4	96.8
Bright-1.1	HDYGNYVDY	7.20	66.2	2.6	98.7
Bright-1.2	DGNYQEDYFDY	5.62	63.1	5.9	98.6
Bright-1.3	EGYAYDVDY	1.91	27.4	23.9	95.6
Bright-1.4	DDYDWYFDV	1.54	59.3	2.8	97.5
Bright-2.1	RGDGNYFFDY	2.57	16.1	14.0	95.0
Bright-2.2	GDEAWFAY	2.27	43.3	6.7	97.1
Bright-2.3	EGDFDY	2.03	14.9	8.1	95.3
Bright-2.4	YYYGSSYFDV	1.84	77.8	0.7	99.2

Antigen	CDRL3	CDR3 Freq (%)	1st V_L Freq (%)[a]	2nd V_L Freq (%)[a]	V_L Homology[b]
OVA-1.1	WQGTHFPLT	11.70	41.4	1.8	92.1
OVA-1.2	QQSNSWYT	4.40	54.5	2.4	94.0
OVA-1.3	QQYSSYPLT	3.38	46.2	1.9	93.9
OVA-1.4	QHHYGTPPWT	2.20	49.7	2.1	93.7
OVA-2.1	WQGTHFPLT	5.32	33.3	2.3	93.7
OVA-2.2	QQYSSYPLT	4.05	43.6	1.1	94.3
OVA-2.3	QQYNSYPLT	3.46	20.1	4.5	92.3
OVA-2.4	QQHYSTPWT	2.01	50.2	2.6	95.3

Table 3 continues to the next page

Table 3 continued from the previous page

Antigen	CDRL3	CDR3 Freq (%)	1st V_L Freq (%)[a]	2nd V_L Freq (%)[a]	V_L Homology[b]
C1s-1.1	WQGTHFPQT	12.95	68.8	1.1	97.9
C1s-1.2	QQWSSYPQLT	6.94	60.9	5.3	90.0
C1s-1.3	QNDHSYPLT	3.81	58.5	3.7	94.5
C1s-1.4	QQGQSYPWT	3.16	70.8	1.1	98.5
C1s-2.1	FQGSHVPLT	17.10	5.7	4.7	90.4
C1s-2.2	QQSNEDPWT	2.62	65.7	2.8	97.4
C1s-2.3	WQGTHFPH	2.20	36.1	18.5	96.5
C1s-2.4	WQGTHFPT	2.15	39.2	15.6	96.9
Bright-1.1	LQYASSPFT	6.64	74.0	1.0	98.3
Bright-1.2	WQGTHFPRT	4.73	60.8	1.5	97.9
Bright-1.3	QQNNEDPRT	4.51	61.8	3.7	97.8
Bright-1.4	QQRSSYPLT	3.59	68.4	0.8	96.5
Bright-2.1	WQGTHFPQT	7.24	44.5	5.7	95.8
Bright-2.2	QQGQSYPWT	4.50	71.3	1.0	98.8
Bright-2.3	LQYASSPYT	3.12	70.7	2.0	98.6
Bright-2.4	FQGSHVPWT	2.58	47.3	3.8	95.0

Table 3: The frequency and homology of highly ranked sequences from the different immunized animals

Notably, the V_H repertoires were quite distinct even among genetically identical littermates immunized with the same antigen on the same day. For mice immunized with C1s or Bright, each mouse developed a distinct and diverse set of abundant CDRH3 sequences (Figure 8). This suggests that each mouse generates its own unique and highly expressed V_H gene repertoire, which may allow for the discovery of a panel of diverse antibodies. One exception however was that in the cohort of OVA-immunized mice we observed that a few abundant CDRH3 sequences also present at high frequency in other mice, suggesting that the corresponding antibodies may be poly-specific. Not

surprisingly, some moderately represented CDRH3 sequences from animals that received adjuvant only, were also present in immunized mice (Figure 8).

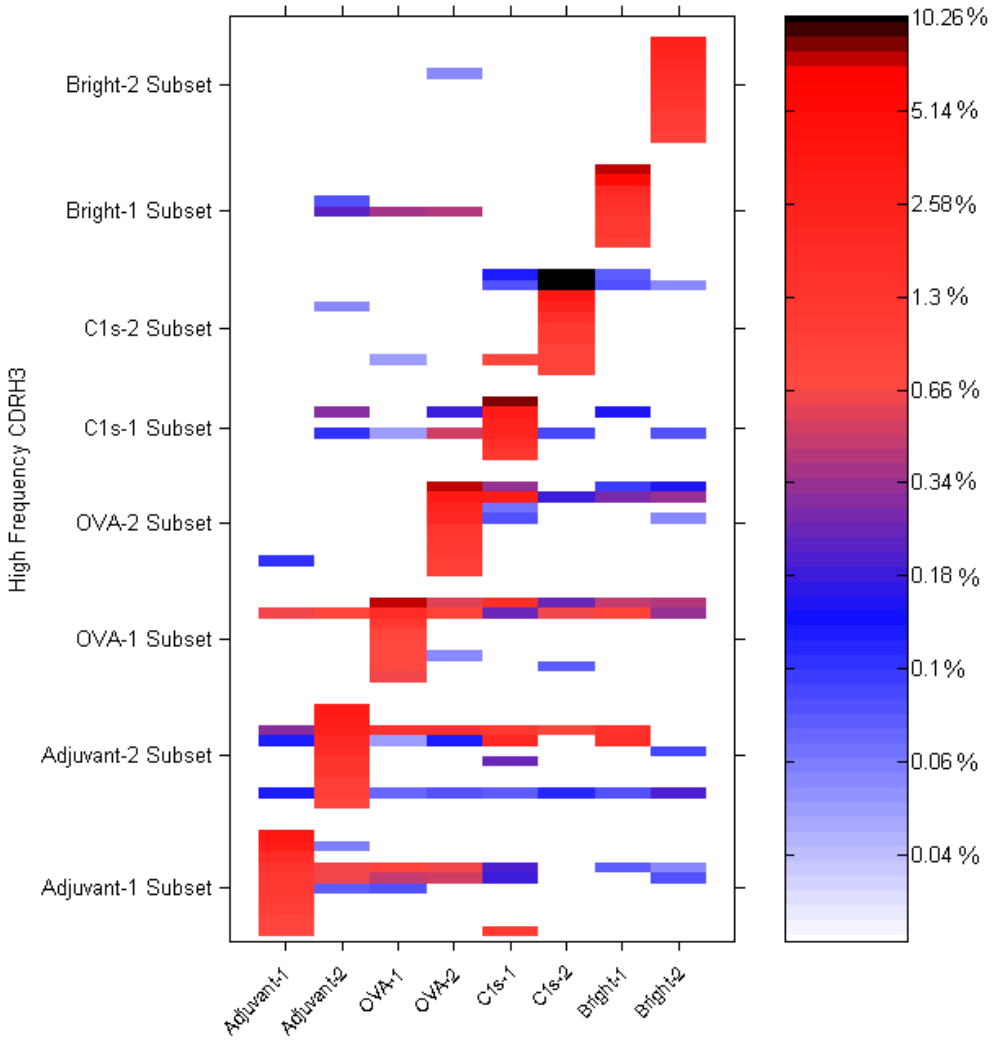

Figure 8: Comparison of high frequency CDRH3s reveals unique V_H genes in each mouse

Antibodies encoding these sequences were probably specific to adjuvant or to common natural antigens. CDRL3 diversity was lower with several promiscuous

sequences represented at high frequency in several mice (data not shown). Fourth, even though the BM-PC V_H repertoires were largely comprised of sequences unique to each mouse, principal component analysis of CDRH3s shared between mice revealed distinct clustering of the data for each cohort (i.e., same cage and litter) immunized at the same time but with different antigens (Figure 9). This signature likely reflects environmental factors, such as the antigenic history of the animal groups, and suggests that V gene repertoire analysis may provide valuable diagnostic information.

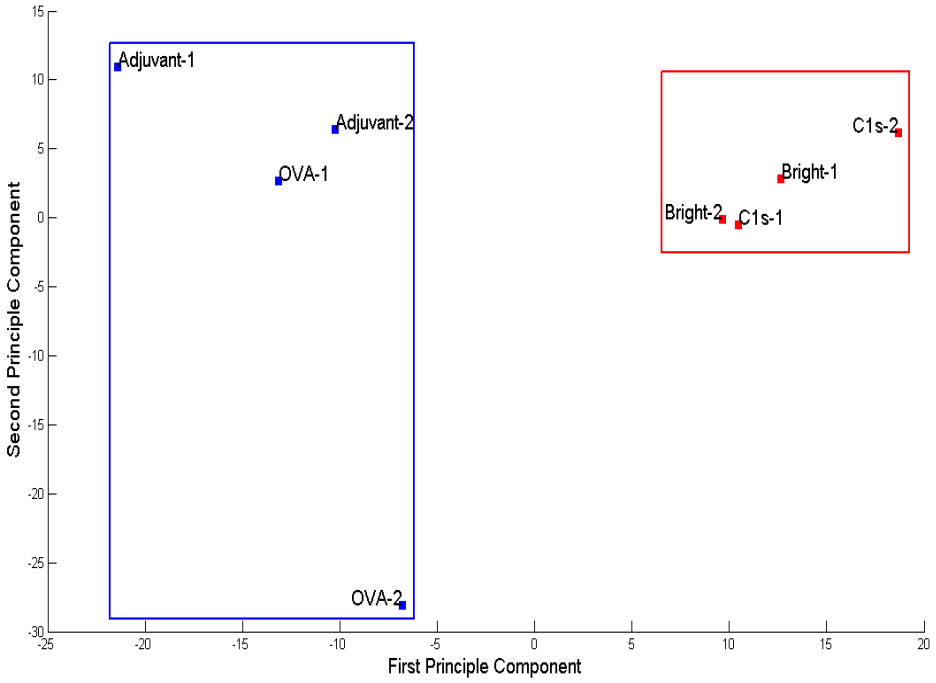

Figure 9: Principal component analysis (PCA) of CDRH3 sequences from the BM-PC repertoires of different mouse groups

It should be noted that a few copies (typically <5) of the most abundant CDRH3 sequences raised to a given antigen were observed at very low levels (typically <0.1%) in the CDRH3 repertoires of mice receiving other antigens. Since several of the respective V genes were shown to encode antigen-specific antibodies, we believe that the presence of these sequences in mice immunized with other antigens might originate from low levels of cross-sample contamination, a conclusion supported by the biased distributions of common CDRH3 sequences within the same cohort. Because of the high sensitivity of 454 DNA sequencing, even with the utmost care it is not possible to completely rule out low-level contamination (sequence noise) during library preparation/multiplex sequencing. Although an important consideration for studies aiming to compare unbiased repertoires [69], [130], sequence noise does not impact the methodology described herein, since the most abundant V genes in the BM-PC repertoire are represented at levels 20- to more than 100-fold higher than the sequence noise level.

Manual screening of small combinatorial libraries of scFvs in *E.coli* using BM-PC V genes led to a low yield of antigen-specific clones (<4 positive clones per 96 well plate, data not shown). Upon further analysis, most of these scFvs displayed low apparent affinity by ELISA and/or poor expression and aggregation. We reasoned that this was a consequence of combinatorial pairing: even if a V_L and a V_H gene are represented at 5% of the cDNA pool, assuming no PCR biases in scFv assembly, the probability of correct pairing is only 0.25%, and therefore discovery of positive clones would require an extensive amount of screening. To overcome these problems, and to avoid screening altogether, we hypothesized that V_L and V_H genes represented at approximately the same abundance likely arise from the same plasma cell and hence, are naturally paired. To test this hypothesis, the top 4-5 most abundant full-length V_L and V_H genes from each mouse

(excluding V_H sequences that were cross-represented in adjuvant-only mice), which accounted for a minimum of 0.5% of the repertoire, were gene synthesized as pairs, recombinantly expressed, and tested for antigen binding. Synthetic genes were constructed by robotically assisted, high-throughput DNA synthesis as described in the materials and methods section. Briefly, gene fragments (lengths from 200 to 500 nucleotides) were generated using inside-out nucleation PCR reactions. The design of these fragments and relevant overlaps was automated using customized software to facilitate robotic synthesis and assembly. Alignment and "padding" of the sequences at either end yielded genes of identical length and permitted the use of a generic overlapping assembly strategy that ensured the greatest oligonucleotide re-use. In this manner, up to 48 V_L and 48 V_H genes could be synthesized and validated for correct ORF by one researcher within one week, at a reagent cost <$2,000. In most cases, V_L and V_H pairing was determined by rank ordering of CDR3 frequency within the repertoire. In cases where two V_L or V_H genes were found at very similar frequencies, we constructed multiple V_L-V_H combinations. Paired V genes were then expressed as scFv fragments in *E.coli*. ELISA analysis of bacterial lysates indicated that the resulting antibodies were overwhelmingly antigen-specific (~78%): we obtained 21/27 antigen specific antibodies from six mice immunized with three different protein antigens (Table 4). To further evaluate the utility of this simple pairing strategy, we constructed a combinatorial library of scFvs comprising the 4 most abundant V_L and V_H genes from each of the two mice immunized with C1s. scFv antibody fragments were expressed in *E.coli*; binding analysis by ELISA revealed that all of the highest antigen-binding clones possessed the same V_L-V_H gene combinations predicted by our pairing strategy (data not shown).

V_L-V_H pair	% V_L	CDRL3	% V_H	CDRH3	scFv binding
α–OVA					
1.1L-1.1H	11.70	WQGTHFPLT	7.11	GSSYYAMDY	+
1.2L-1.2H	4.40	QQYNSYPLT	1.10	LLWLYAMDY	+
1.3L-1.3H	3.38	QQSNSWYT	0.57	DVYDGYAMDY	+
1.4L-1.4H	2.20	QHHYGTPPWT	0.54	NPYAMDY	-
2.1L-2.1H	5.32	WQGTHFPLT	7.61	RTTVSRDWYFDV	+
2.2L-2.2H	4.05	QQYNSYPLT	3.23	YYYGSSAMDY	+
2.3L-2.3H	3.46	QQYSSYPLT	2.22	DGWYYFDY	+
2.4L-2.4H	2.01	QQHYSTPWT	2.10	EDDYDLFAY	+
α–C1s					
1.1L-1.1H	12.95	WQGTHFPQT	7.93	GNYYYAMDY	+
1.2L-1.1H	6.94	QQWSSYPQLT	7.93	GNYYYAMDY	+
1.3L-1.2H	3.81	QNDHSYPLT	2.64	DMISYWYFDV	+
1.4L-1.3H	3.16	QQGQSYPFT	1.67	EDYGNYWYFDV	+
1.4L-1.4H	3.16	QQGQSYPFT	1.67	EGYYYGSSYFDY	-
2.1L-2.1HA	17.10	FQGSHVPLT	10.99	SDRYDGYFDY	+
2.1L-2.1HB	17.10	FQGSHVPLT	9.93	SDRFDGYFDY	+
2.2L-2.2H	2.62	QQSNEDPWT	3.30	WLLLAY	+
2.3L-2.2H	2.20	WQGTHFPH	3.30	WLLLAY	+
2.3L-2.3H	2.20	WQGTHFPH	1.65	SDGYYYFDY	+
2.4L-2.4H	1.64	QQHYSTPFT	1.15	YYDYDKAYYFDY	-
α–Br					
1.1L-1.1H	6.64	LQYASSPFT	7.20	HDYGNYVDY	+
1.2L-1.2H	4.73	WQGTHFPRT	5.62	DGNYQEDYFDY	-
1.3L-1.3H	4.51	QQNNEDPRT	1.91	EGYAYDVDY	+
1.4L-1.4H	3.59	QQRSSYPLT	1.20	YDYGKDFDY	+
2.1L-2.1H	7.24	WQGTHFPQT	2.57	RGDGNYFFDY	+
2.2L-2.2H	4.50	QQGQSYPWT	2.27	GDEAWFAY	-
2.3L-2.3H	3.12	LQYASSPYT	2.03	EGDFDY	-
2.4L-2.4H	2.58	FQGSHVPWT	1.63	GGNYDYAMDY	+

Table 4: Antigen binding of antibody single-chain variable fragments (scFvs) from high frequency V_L and V_H genes

Mouse C1s-2 displayed the highest serum titers and therefore, antibodies from this mouse were selected for biophysical characterization of antigen binding affinity by surface plasmon resonance (Biacore; data shown in subsequent paragraph) and sandwich ELISA (Figure 10).

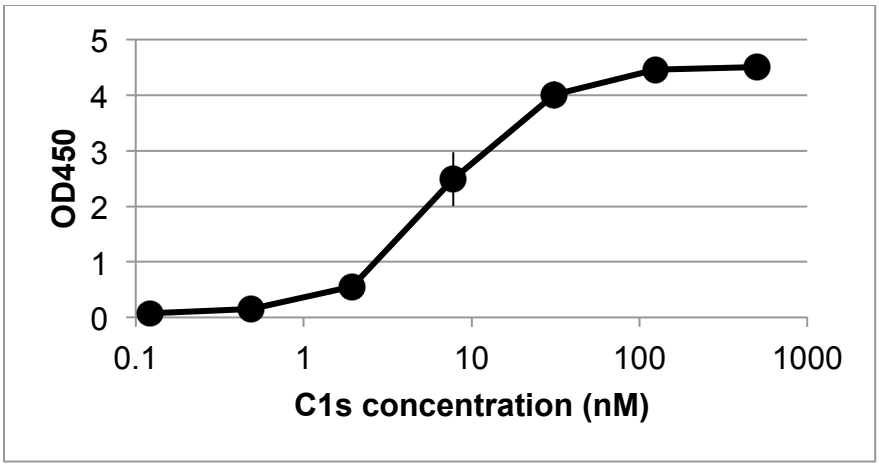

Figure 10: Sandwich ELISA by coated synthetic anti-C1s scAb 2.1L-2.1H-B capturing with C1s and detecting with characterized anti-C1s high-binder full-length IgG

Antibodies were recombinantly expressed and purified as monomeric scFv fragments in *E.coli* and as full-length IgG antibodies in HEK 293F cells. Pairing of the most abundant light (2.1L) and heavy (2.1H-B) V genes (17.10% and 9.93% CDRL3 and CDRH3 frequencies, respectively) from mouse C1s-2 yielded an antibody with a K_D of 20 nM as a scFv (k_{on}=2.3 x 10^4 M^{-1} sec^{-1}; k_{off}=5.0 x 10^{-4} sec^{-1}) and unexpectedly, a slightly lower monovalent K_D of 50 nM (k_{on}=2.4 x 10^4 M^{-1} sec^{-1}; k_{off}=1.2 x 10^{-3} sec^{-1}) as an IgG. From the same mouse, pairing of C1s 2.2L with 2.2H (2.62% and 3.30% CDRL3 and CDRH3 frequencies, respectively) resulted in an IgG that displayed low binding affinity (K_D of ~500 nM, data not shown). However, the pairing of C1s 2.3L with 2.2H (2.20%

and 3.30% CDRL3 and CDRH3 frequency, respectively) yielded an IgG with sub-nanomolar binding affinity (K_D=0.43 nM, k_{on}=4.5 x 10^5 M^{-1} sec^{-1}; k_{off}=1.9 x 10^{-4} sec^{-1}, indicating that the natural pairing is likely 2.3L-2.2H (Table 5). Not surprisingly, the antibodies were suitable for functional assays, such as sandwich ELISA and immunoprecipitation of C1s from human serum as shown above.

V_L-V_H pair	2.1L-2.1HB (scAb)	2.1L-2.1HB (IgG)	2.3L-2.2H (IgG)
% V_L	17.10	17.10	2.20
% V_H	9.93	9.93	3.30
CDRL3	FQGSHVPLT	FQGSHVPLT	WQGTHFPH
CDRH3	SDRFDGYFDY	SDRFDGYFDY	WLLLAY
K_{on}	2.35 x 10^4	2.38 x 10^4	4.51 x 10^5
K_{off}	4.98 x10^{-4}	1.21 x 10^{-3}	1.92 x 10^{-4}
K_D (nM)	20	50	0.43

Table 5: Biophysical characterization of the different format of anti-C1s molecules derived from the BM-PC repertoires of mouse C1s-2

DISCUSSION

Our approach capitalizes on the mining of the repertoire of BM-PCs, a population of B cells that is responsible for the synthesis of the large majority of circulating immunoglobulins in animals [123]. While we have validated this methodology in mice there is no reason to believe that the same approach cannot be readily extended to primates including humans. Furthermore, it is possible that this technology could be extended for antibody discovery with more complex antigens such as viral and bacterial pathogens, as in these situations BM-PCs may still develop polarity to a single antigen. The mechanisms that dictate the selection of B cell differentiation into plasma cells and homing into the bone marrow are complex and appear to partially relate to high antigen

affinity [50], [132]. Since the highly abundant BM-PCs correspond to abundant circulating antibodies, it is plausible to hypothesize that these antibodies have been selected by the immune system (at least partly) because they display more potent pathogen neutralization. Therefore antibodies generated by the mining of the BM-PC repertoire may prove particularly useful for therapeutic purposes. The hybridoma technology and other B cell immortalization methods interrogate the antibody producing cells in pre-PC B cell populations, specifically in memory B cells, or in circulating short lived plasmablasts [112]; while, fully differentiated plasma cells are not amenable to most of these analyses since they do not survive outside their biological niches. Very recently, Jin *et al*. used microwell arrays and single cell cloning to isolate antibodies from spleen plasma cells [119]; however, despite the use of a sophisticated screening technology small numbers of antigen-specific clones could be isolated and consequently information on the repertoire and relative abundance of V genes could not be obtained by this method.

Here we report on a simple, rapid, and fundamentally new approach for antibody isolation without screening that capitalizes on the mining of BM-PC antibody repertoires. The ability to take advantage of high-throughput DNA sequencing, bioinformatics analysis, and automated gene synthesis can lead to the isolation and expression of mAbs with minimal effort. In our hands, we have estimated that it takes about 10 man-hours for sample preparation for DNA sequencing. With automated bioinformatics processing of the 454 sequencing data, no extra effort is required to identify highly abundant V_L and V_H genes for DNA synthesis. Synthetic genes can be constructed either by an automated facility (as described herein) or through commercial gene synthesis vendors (such as IDT's gBlock). Furthermore antibody genes can be codon optimized as desired for either

bacterial or mammalian expression and subsequent characterization studies. Thus, in terms of effort by dedicated personnel (not including DNA sequencing and synthesis, which are carried out by multi-user services) and timeline required for antibody discovery, our method compares very favorably to hybridomas, B cell immortalization, and B cell screening/single cell cloning methodologies. Currently, the most expensive part of our antibody discovery process is DNA sequencing followed by gene synthesis; however the cost for these technologies are declining at a rapid and exponential pace, resembling Moore's law for microelectronics [76], [133]. Taken within this context, the expense for our approach to antibody discovery will eventually not be a limitation.

Chapter 3: Intrinsic Bias and Public Rearrangements in the Human Immunoglobulin Vλ Light Chain Repertoire

This chapter is reproduced with minor modifications from its initial publication:
K. H. Hoi and G. C. Ippolito, "Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire," *Genes Immun*, vol. 14, no. 4, pp. 271–276, Jun. 2013.

Acknowledgements:
The author would like to acknowledge Professor Gregory Ippolito for assistance in writing and editing this chapter and the associated manuscript. Also, the author would like to acknowledge Jaime O'Neal and Chhaya Das for their technical assistance. We would also like to acknowledge Arvind Rajpal, Tracy Kuo, and Marina Sirota for providing the Rinat-Pfizer Twin sequencing data and collegiality.

INTRODUCTION

Immunoglobulin B cell receptor (BCR) diversity is an essential component of adaptive immunity for the recognition of a diverse constellation of foreign antigens. As a heterodimeric protein containing two heavy-chains and two light-chains, immunoglobulin is encoded by the rearrangement of variable (V), diversity (D), and joining (J) gene segments located in the immunoglobulin heavy locus (IGH) on chromosome 14 whereas the light chain is encoded by V-gene and J-gene rearrangements derived from either the immunoglobulin kappa locus (IGK) on chromosome 2 or the lambda locus (IGL) on chromosome 22. Genetic processes are a critical and fundamental step in the generation of extensive BCR diversity [29], [134]–[136]. A diversity of 1×10^{11} unique BCRs can be achieved in theory via random V(D)J recombination together with pairing of heavy and light chains. Additionally, due to imprecision in the V(D)J joining process itself, as well as enzyme-catalyzed reactions which can result in the addition of nontemplated (N/P) nucleotides and the introduction of somatic hypermutations (SHM), these mechanisms

together further expand the potential diversity of the immunoglobulin repertoire to more than one peta-BCRs ($1x10^{15}$) [36]. Contrarily, the actual diversity in humans is likely to be several orders of magnitude lower, due in part to physiological limitations such as the number of B cells which can be derived during the finite lifespan of the organism. Actual diversity has been experimentally estimated to be on the order of 1-10 million [69]–[71]; hence, it seems implausible that random processes should be a primary constituent for BCR diversity, and thus selective mechanisms, be it genetic and/or somatic, must be in place to maintain the observed range of BCR diversity.

In support of this view, characterization of heavy and light chain immune repertoires has revealed preferential, nonrandom usage of particular IGH genes [70], [137]–[141] as well as IGK genes [142]–[145]. One recent study of the human IGK repertoire [145], using next-generation deep sequencing of more than 60,000 IGK cDNA reads generated from peripheral blood B cells derived from four ethnically different individuals, showed a significantly biased use for particular IGK variable (IGKV) and IGK joining (IGKJ) genes; interestingly, the authors also reported that a surprisingly high percentage (up to 60.2%) of IGK protein sequences were shared, or so-called "public", between any two of the four individuals examined, leading to the conclusion that the repertoire of unique κ chains is merely on the order of thousands as compared to prior theoretical repertoire estimates comprising 10^5 – 10^6 different IGK genes. This observation of frequent public sequences in the IGK repertoire differs from previous reports [142]–[144], presumably because at most only a total of a few hundred cells per individual were sequenced, and probably reflects the greater sensitivity and coverage made possible by the deep-sequencing techniques used in this study. Furthermore, the authors speculated that the existence of public IGK sequences might reflect: (i) a limited

number of rearrangements that do not introduce conformational clashes upon pairing with the heavy chain thus resulting in productive pairings; (ii) selection of light chains that do not interfere with antigen recognition, the latter being mediated near-exclusively by the IGH; or lastly, (iii) that certain IGK rearrangements are either counter-selected due to potential auto-reactivity or are positively selected as a response mechanism to superantigen-like molecules on microbes [145].

In contrast with the more common studies of IGH and IGK repertoires, surprisingly little has been published regarding the IGL repertoire, whether in humans or various humanized mouse models [146]–[151]. Thirty-five functional IGLV segments have been classified within 10 gene families and 4 functional IGLJ segments [66], [152]–[154]; CDR-L3 length is highly restricted while N/P nucleotide additions on average add only one new codon to the CDR-L3 [147], [149]. One prior theoretical calculation of the combined IGK plus IGL CDR3 repertoire has been estimated to be on the order of >1.6 x 10^5 [70]. The CDR-L3 repertoire, specifically, has been estimated to comprise perhaps as many as 338,130 unique sequences [155]. Regarding its expression, preferential IGLV gene usage has been observed in humans as well as in transgenic mice engrafted with human lambda loci [146], [148]. Furthermore, Richl et al. [151] also reported preferential IGLV gene usage in human neonatal and adult B cells through the analysis of 236 productive IGL sequences. These results suggested that the underlying gene recombination processes might be non-random yielding a preferred IGL repertoire.

Here we have analyzed the IGL repertoire at great depth using next-generation sequencing (NGS) data from peripheral blood B cells obtained from two volunteers and compared it with the IGL repertoires derived from splenic B cells in NOD-*scid-IL2Rγ*^{null}

mice (humanized mice) engrafted with human umbilical cord blood stem cells [18]. For additional comparison, deep sequencing data obtained independently by another research group and derived from two pairs of identical twins [138] were included in our analysis. As noted in the NGS study of the human IGK repertoire by Jackson et al. [145], we report here the frequent occurrence of public rearrangements within human and humanized mice IGL repertoires. Interestingly, public rearrangements include not only identical CDR-L3 peptide sequences but also identical full-length IGL protein sequences. Furthermore, we report that notable differences exist between public and private CDR-L3s in terms of the degree of N/P nucleotide addition, somatic hypermutation, IGLV1 gene family usage, and amino acid utilization.

MATERIALS AND METHODS

Humanized mice

NOD.Cg-Prkdc^{scid} IL2rg^{tm1Wjl} (NOD-scid-IL2Rγ^{null}) mice were obtained from the Jackson Laboratory (Bar Harbor, ME). NOD-scid-IL2Rγ^{null} mice were irradiated with 100 cGy of gamma irradiation at 1-2 days of age. Upon irradiation, 50 μL of 3 x 10^4 human CD34$^+$ hematopoietic cells derived from umbilical cord blood were injected via intracardiac route (Lonza, Cat. 2C-101B). HuMs1 and HuMs2 were both engrafted with the same donor, while HuMs3 was engrafted with a different donor. The humanized mice were housed at the Animal Resource Center of the University of Texas at Austin. All experiments were conducted following the guidelines provided by the university's Institutional Animal Care and Use Committee (protocol number AUP-2010-00089).

B cell preparation, RNA extraction, cDNA generation, and PCR amplification

Detailed procedures have been described previously [18]. Briefly, mononuclear cells from two anonymous healthy donors, (both healthy females with one in her 30s and the other in her 50s), were prepared from whole blood using Histopaque-1077 density centrifugation and humanized mice spleens were harvested post euthanization. These single cell suspensions were used for total RNA preparations, and Oligo-dT cDNA was generated which was then used as template for an optimized PCR amplification of Vλ-Jλ recombinants. Samples were collected and prepared independently over a period of four months.

Next generation sequencing of IGL repertoires

Detailed procedures have been described previously [18]. Briefly, samples were serially acquired and PCR amplicons were gel-purified and prepared independently over a period of four months. Samples were then submitted to the University of Texas Genome Sequencing and Analysis Facility for library construction and Roche GS-FLX 454 high-throughput sequencing.

Bioinformatics analysis

IGL repertoire sequences were obtained from two adults' peripheral blood mononuclear cells (PBMCs), from splenic B cells isolated from three humanized mice, and from a pool of immature B cells isolated from three humanized mice [18]. Total RNA was isolated from the samples and library was generated with previously reported sets of primers [83]. Sequences were filtered to meet a minimum of 350 bps length requirements

and each had to contain an in-frame CDR-L3 region. These filtered sequences were submitted to IMGT/HighV-QUEST [89] for complete annotations, namely, Vλ and Jλ gene segment assignment, N- and P-nucleotide addition, and somatic hypermutation. Annotated information was parsed and extracted using Perl scripts for the analysis described in this manuscript. Public CDR-L3 was defined as CDR-L3 peptide sequences that can be found in two or more samples among the ten samples we examined. Only unique IGL nucleotide sequences were analyzed to avoid over-counting of duplicated sequences. All raw 454 sequences have been deposited to the NCBI Sequence Read Archive (Accession SRA049345).

Rinat-Pfizer Twin Sequences

Sequences from two pairs of adult monozygotic twins were kindly provided by the Rinat-Pfizer scientists [138]. Twin pair A [denoted in this manuscript as A1, A2] is of Western European and Mediterranean descent at the age of 54 years old. Twin pair B [denoted in this manuscript as B1, B2] is of Ashkenazi Jewish descent at the age of 57 years old. Twin pair B was diagnosed with Multiple Sclerosis. These sequences were subjected to the same sequence analysis processing procedures and bioinformatics analysis as described in the bioinformatics analysis section of this manuscript.

Statistical methods

GraphPad Prism version 5.00 for windows was used to conduct statistical analysis. ANOVA (Kruskal-Wallis test) and Mann-Whitney test were used to determine the statistical significance of the population means.

RESULTS

B cells were sorted and IGL cDNAs were sequenced using Roche GS-FLX 454 technology, as described previously [18]. Additionally, independently derived IGL cDNA sequences from two sets of identical twins were provided by the Rinat-Pfizer group [138] to corroborate our analysis. Only DNA sequences with more than 350bps and identified by IMGT/V-QUEST as "productive" were considered for further analysis. "Public" CDR-L3s were defined as CDR-L3 peptide sequences present in two or more samples whereas "private" CDR-L3s were unique to their single respective sample. In the work previously published by our group [18], under the same sample preparation and sequence analysis procedures as in this report, we found less than 0.001% of CDR-H3s to be shared among the human and humanized samples, which is consistent with the findings in other NGS studies [69], [71]. On the other hand, using the dataset derived from the same procedures, we analyzed the overlap of CDR-K3 between the human samples and found that about 19.7% of CDR-K3s were shared, which is similar to, although 3-fold less, than the degree of public sharing reported by Jackson et al. [145]. The frequency of public CDR3s found in our IGH library (<0.001%) and in our IGK library (~20%) agree with prior publications and, therefore, seem to indicate a minimal-to-zero introduction of sequencing artifacts and, furthermore, that our sample preparation and sequence analysis procedures faithfully preserved the integrity of the repertoires. Any artifacts which could have been introduced would have been expected to appear disproportionately, but this was not observed in the previous work nor in the CDR-L3 dataset analyzed here.

Analysis of the human samples (HuPBC1, HuPBC2), humanized mice samples (HuMs1, HuMs2, HuMs3, HuMsImmB), and Rinat-Pfizer twin samples (A1, A2, B1, B2), revealed that any given individual sample shared at least 20% of its CDR-L3s with

at least one other of the nine samples (Figure 11). IGL repertoires from humanized mice on average contained higher percentages of public CDR-L3s compared to the human counterparts (~57% versus ~30%) possibly due to an expanded compartment of naïve B cells in the humanized mice [18].

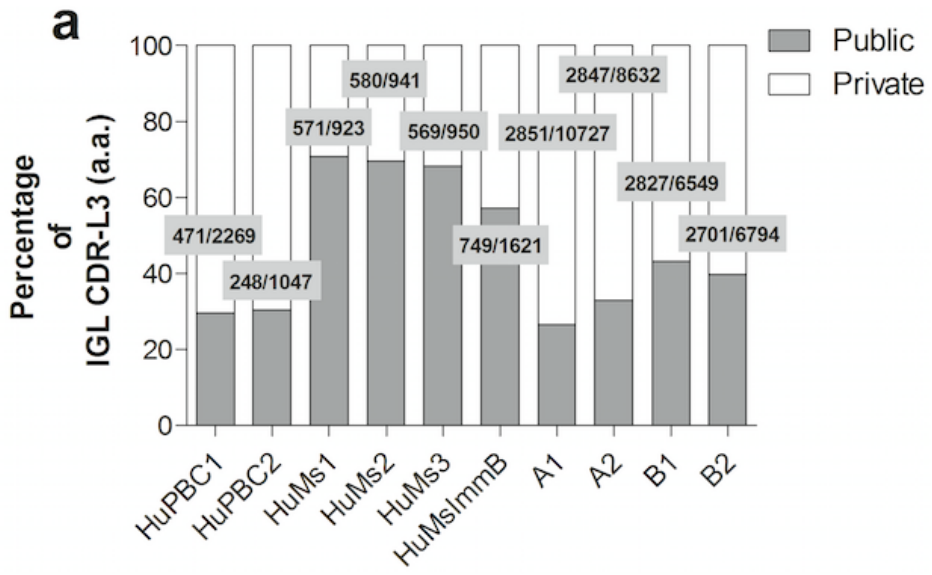

Figure 11: Percentage of public IGL CDR-L3s across all samples

The human peripheral B-cell samples HuPBC1 and HuPBC2 (Figure 12b) exhibited a lower percentage of public CDR-L3s as compared to the complete comparison across all samples. The majority of the CDR-L3s in the two human samples were therefore distinct and hence private. Additionally, Rinat-Pfizer twin samples also indicated a lower percentage of public CDR-L3s as shown in Figure 12c. Both of these human datasets suggest that a greater number of naïve CDR-L3s were displaced to give rise to more "customized" CDR-L3s to cope with increased exposure to pathogenic

environments in both the human groups which was diminished in the specific pathogen-free environment of the humanized mice.

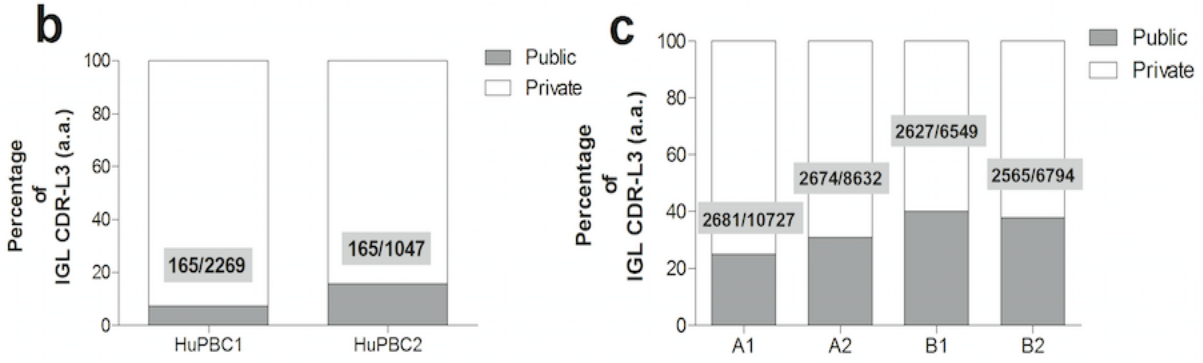

Figure 12: b-Percentage of public IGL CDR-L3s across in-house human samples c-Percentage of public IGL CDR-L3s across Rinat-Pfizer twin samples

On the other hand, as shown in Figure 13, the reverse was observed in the humanized mice samples where at least 40% of CDR-L3s were public. Nevertheless, frequent public CDR-L3s were still observed across all samples exposed to presumably different pathogenic environments; this includes public CDR-L3 repertoires shared between the two independently derived human datasets (Figure 14).

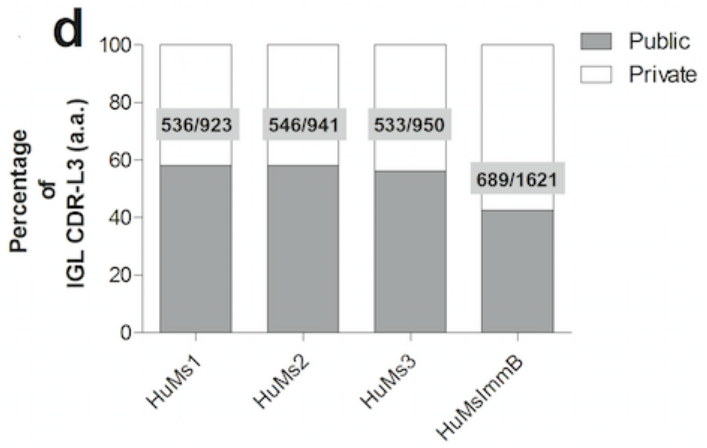

Figure 13: Percentage of public IGL CDR-L3s across all humanized mice samples

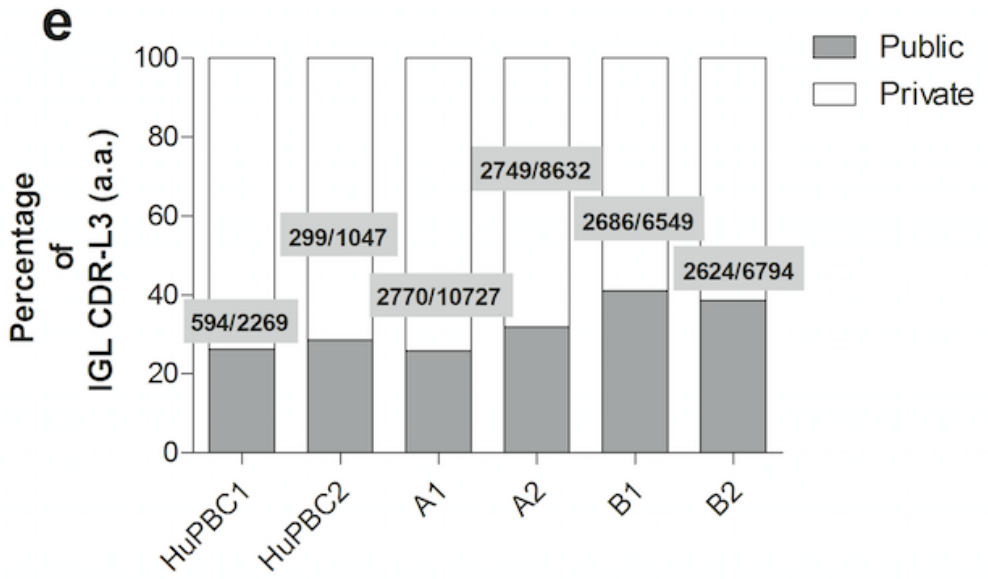

Figure 14: Percentage of public IGL CDR-L3s across all human samples

Additionally, these findings imply that preferential IGL gene rearrangements occur regardless of the repertoire's developmental microenvironment. Therefore, there is a strong indication that preferential λ–chain gene rearrangement could be due to intrinsic genetic mechanisms independent of selective forces exerted at the cellular level. It is still unclear, however, as to the true cause of the difference in public CDR-L3 percentages between the human and the humanized mice samples. It is worthwhile to note that humanized mice tend to have impaired immune responses which suggests that many clones might not be significantly hypermutated or affinity-matured.

Next, we separated full-length nucleotide sequences containing public or private CDR-L3 into individual FASTA files. These public/private-categorized sequences were

submitted independently to IMGT to ascertain N/P nucleotide addition, somatic hypermutation (SHM), CDR-L3 peptide sequence, V-J gene utilization, and full-length IGL protein sequence for each public/private group. Subsequently, we characterized the public and private IGL sequences in terms of V-J gene segment utilization and we found a decrease in the use of IGLV1 gene family in the public CDRL-3 sequences for the humanized mice group (Figure 15).

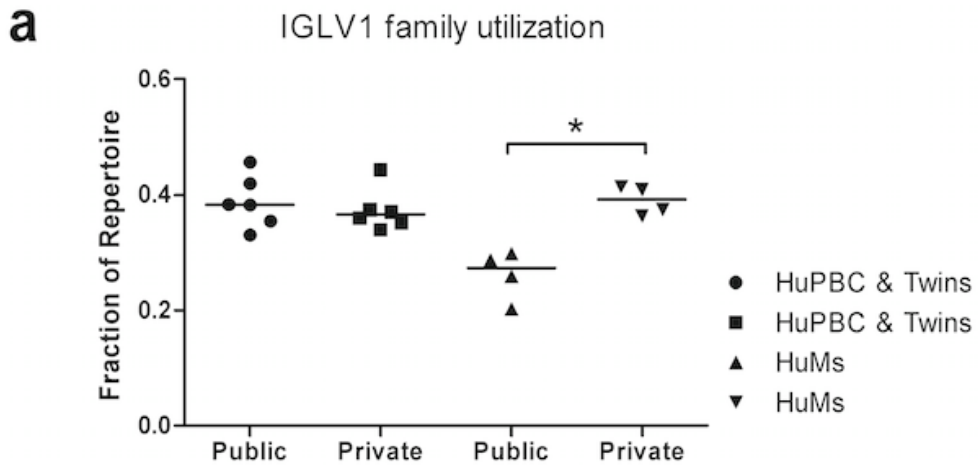

Figure 15: IGLV1 family repertoire usage comparison

This difference was not observed in the human group. Utilization of the remaining V and J gene segments did not show statistically significant differences across samples. CDR-L3 N/P nucleotide addition and IGL SHM were also examined, and the value for each sample was calculated as an average of all the SHM counts for the IGL sequences (FR1-FR4) derived from each sample. In terms of N/P nucleotide addition (Figure 16), there was a statistically significant difference between the public CDR-L3 sequences and

the private CDR-L3 sequences ($p<0.001$). In general, the private sequences demonstrated approximately two-fold more N/P nucleotide additions than the public CDR-L3 sequences, indicating how preferential gene rearrangements are favored by fewer N/P additions, resulting in less diversity, enabled by unaltered joining of a germline IGLV gene segment to an IGLJ gene segment due perhaps in part to microhomology-mediated joining of the gene segments.

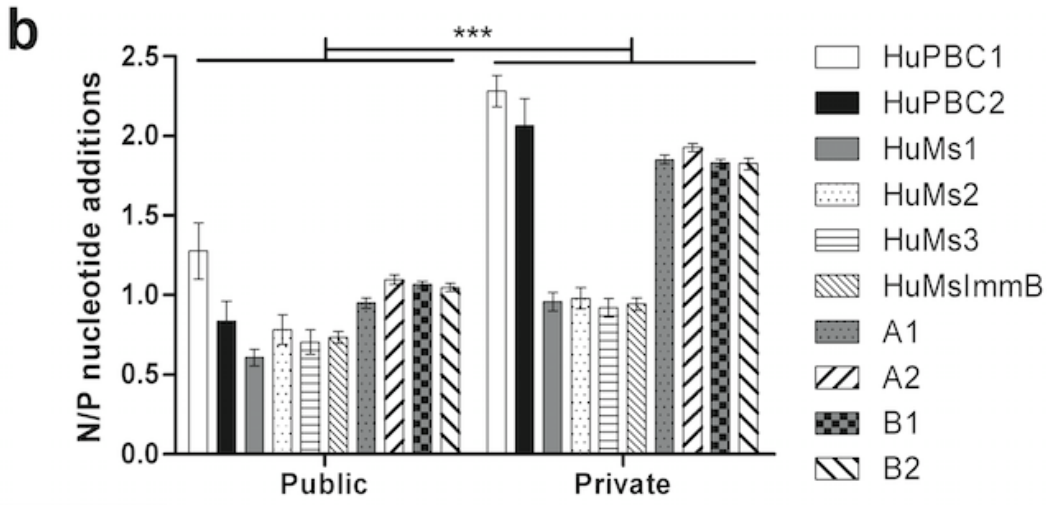

Figure 16: N/P nucleotide additions comparison

As a result, the likelihood of such unaltered CDR-L3 sequences occurring among individuals appears to be higher; however, it would require more evidence to conclude that the direct cause for public CDR-L3s can be solely attributed to fewer N/P nucleotide additions. Nevertheless, it was an unexpected observation that overall N/P nucleotide addition would be suppressed in the public CDR-L3 group as compared to the private CDR-L3 group.

On the other hand, in Figure 17, no significant difference in SHM was observed between the public and private CDR-L3 groups for the humanized mice samples; however, the SHM difference between the public and private groups for the human samples, both the HuPBC and Rinat-Pfizer twin samples, were statistically significant ($p<0.001$).

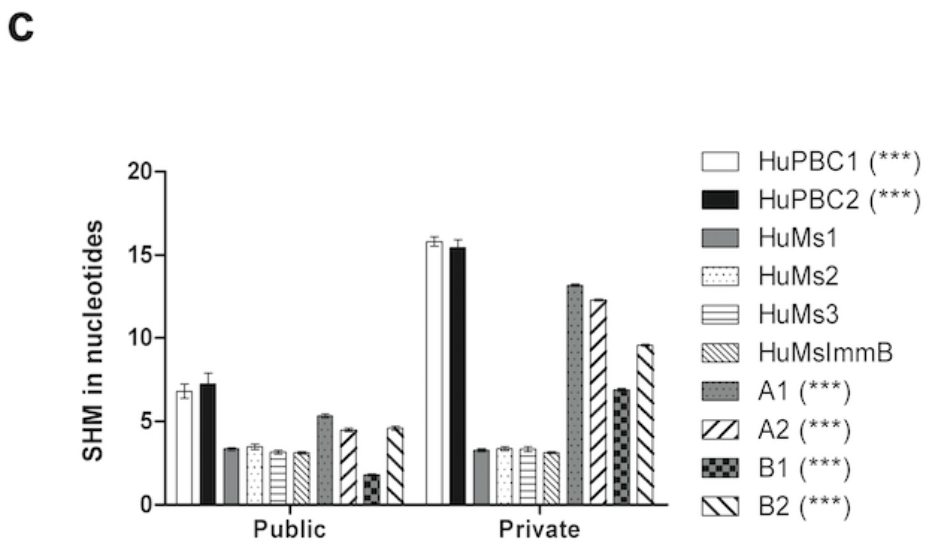

Figure 17: Nucleotide SHM comparison

This difference could be due to the effects of antigenic selection augmenting the SHM numbers for the human samples. Since the humanized mice were housed in a specific pathogen-free environment and that no overt antigenic pressure was applied, germinal center reactions and hence SHM in the humanized mice should have been scant to nonexistent, as has been described generally for this mouse strain. This alone might explain the similarity of SHM observed in both the public and private humanized mice CDR-L3 groups. The significant difference between the public and private groups of

CDR-L3s in the human samples, on the other hand, might reflect extensive past immune responses and the initiation of germinal center reactions, SHM, and selective mechanisms exerted upon antigen-specific B cells.

We investigated the amino acid composition of length=11 CDR-L3s which were the most frequently observed across samples. We noticed, in general, that amino acid composition was similar between the public and private groups, except for arginine, proline, tryptophan, serine, and valine usage ($p<0.001$) as shown in Figure 18.

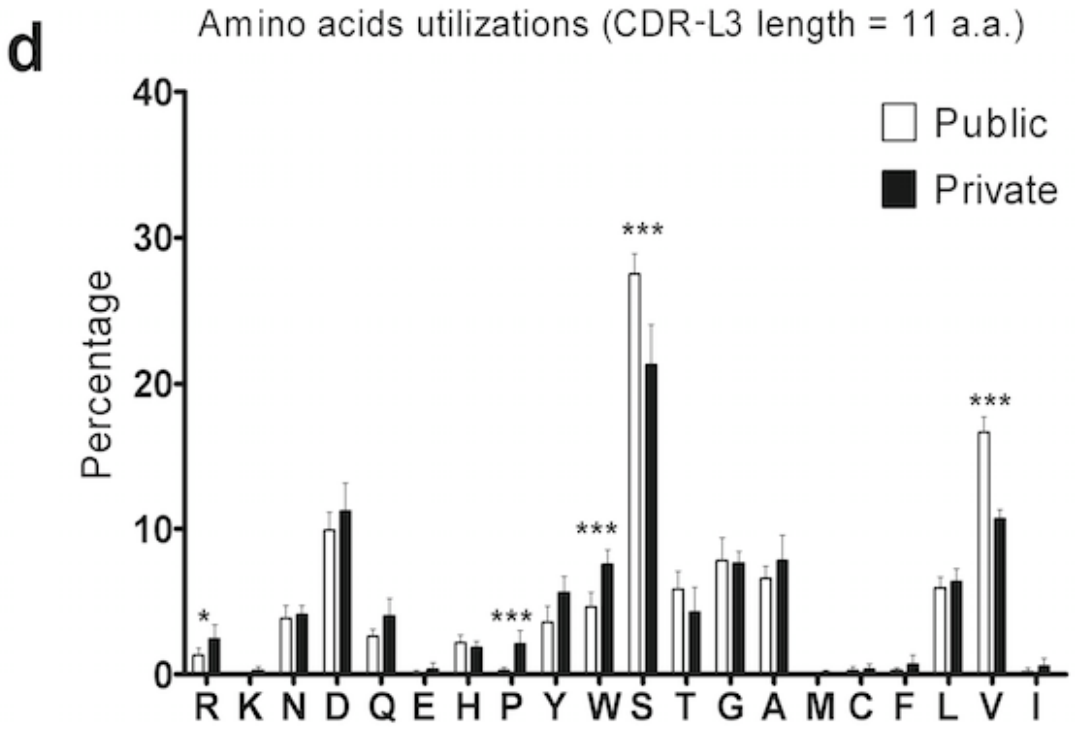

Figure 18: Amino acids utilization for CDR-L3 at length 11

Among those differences, the private groups utilized arginine, proline, and tryptophan ~1%-3% more than within the public groups. On the other hand, the public groups utilized ~6% more serine and valine. Serine was generally abundant in CDR-L3s across all samples overall, due primarily to the inherent serine-rich nature of IGL germline genes. The reduction of serine usage in the private group that we observed here could simply be due to the "customization" of the germline genes after various degrees of antigenic exposures. The enhanced use of arginine, proline, and tryptophan in the private group further support the notion of CDR-L3 "customization" as a result of antigenic challenge since these amino acids tend to promote recognition of pathogenic moieties.

Lastly, upon further investigation of all the public CDR-L3 sequences, we observed that ~10.14% were part of otherwise identical full-length IGL protein sequences (Table 6). As noted in one prior study [150] where a frequency of ~15% identical IGL chains was observed within any single biologic sample assayed (fetal tissue or adult blood), the ~10% frequency in the present study is highly comparable—but striking in its discovery across a multiplicity of independently derived samples.

Public CDR-L3 peptides appended with IMGT Vλ and Jλ assignments	Count	%
Total	4169	-
Non-exact match of CDRL-3+IMGT Vλ and Jλ assignments	1801	43.20
Exact match of CDRL-3+IMGT Vλ and Jλ assignments	2368	56.80

Full-length IGL protein sequences from the public CDR-L3 group	Count	%
Total	26013	-
Non-exact match of the Full-length IGL protein sequences	23375	89.86
Exact match of the Full-length IGL protein sequences	2638	10.14

Top panel: IMGT Vλ and Jλ assignments along with CDR-L3 were considered among the public CDR-L3 group
Bottom panel: Full-length IGL protein sequences were considered among the public CDR-L3 group

Table 6: Enumeration of public CDR-L3s in all samples

DISCUSSION

Public CDR-L3s across individuals can be observed in IGL repertoires similar to that reported for IGK repertoires [145] but in stark contrast to the unique diversity and "privacy" of IGH repertoires [18], [69], [71]. Public CDR-K3s do occur at a very high frequency [145] and at a rate surpassing what we have observed for CDR-L3s. The IGL preferential gene rearrangements we observed do not seem to be affected by the microenvironment suggesting that intrinsic genetic mechanisms of recombination may contribute to this phenomenon. An associated phenomenon in the public CDR-L3 group was found to be suppressed N/P nucleotide addition that could potentially increase the likelihood of public CDR-L3s due to microhomology-mediated, unaltered joining of IGLV and IGLJ gene segments. As a result, the likelihood of public CDR-L3 would be increased. On the other hand, SHM trends observed in our study were inconsistent and did not indicate a clear correlate between SHM and the occurrence of public CDR-L3s. Nonetheless, a high percentage of public CDR-L3s having identical full-length IGL

protein sequences seems to support the naïve newly-formed nature of the IGL public repertoires. That naïve peripheral B cells were specifically sorted in the humanized mice, and that these samples exhibit the most extensively public CDR-L3 repertoire, further supports the notion that it is the primary IGL repertoire which can be publicly shared.

Our findings indicate that the IGL repertoire is significantly constrained in the expression of CDR-L3 peptide sequence and suggests that a considerable fraction of these CDR-L3s is evolutionarily conserved and is expressed "publicly" by the species. By extension, limited λ-chain diversity implies that the combinatorial pairing of IGL light chains with IGH heavy chains might also be constrained. In most resolved antibody structures, the CDR3 loops of IGH and IGL are nearly always in contact at or near the center of the antibody binding pocket [156], [157]. Furthermore, unlike the considerable structural variation and sequence diversity of the CDR-H3, CDR-L3s adopt only a very limited set of distinct "canonical" conformations [14], [158]. The sequence restriction and public sharing of CDR-L3 protein sequences we report here seemingly correlates with the previously appreciated restriction of canonical structures. Even though one previous report did suggest the role of intrinsic genetic factors in the generation of the human IGL repertoire and, moreover, proposed this as a chief mechanism for ensuring the overrepresentation of particular IGLV segments [148], the extent of sequence analysis was insufficient to fully extrapolate their observation specifically to the CDR-L3, as we have done here using NGS deep-sequencing technology.

Previous studies from the Lipsky laboratory using single-cell analysis of IGL repertoires in human fetal and adult tissues have suggested the positive selection of immunoglobulin light chain which is independent of the immunoglobulin heavy chain

[149], [150]. Although based upon a limited number (~100) of single-cell-derived sequences, the repeated occurrence of identical IGL chains was firmly established; however, these identical IGL chains were observed only within any one single sample assayed, contrary to our observation of public occurrences across multiple samples, which we attribute to marked differences in the depth of sequencing coverage between their study and ours. Moreover, these prior single-cell studies documented how identical IGL chains could pair with multiple IGH chains [150]. A more recent analysis by Weigert and colleagues indirectly supports this same observation and similarly concludes the existence of an IGH-independent mode of IGL selection [159]. Using microarray profiling rather than deep sequencing, a set of IGL genes was observed to be uniformly highly expressed without IGH chain restriction (i.e., IGH exhibiting a normally distributed CDR-H3 length spectratype), which they conclude is consistent with the positive selection of particular IGL chains, rather than clonal selection, independent of the IGH chain. Their results suggest that multiple overexpressed IGL chains might pair promiscuously with IGH chains. Indirectly, these results imply the possibility that some of the highly overexpressed IGL chains are in fact identical.

It might be proposed, therefore, that certain CDR-L3s are optimal for binding a range of antigens and therefore have been selected and maintained throughout human evolution. An immediate corollary to this proposal is that certain classes of public CDR-L3s pair preferentially with IGH chains. To address these possibilities and to gain further insight into the human antibody repertoire, we have developed a high-throughput methodology for NGS deep sequencing of natively paired H-L immunoglobulin chains isolated from single B lymphocytes [20]. The nature of this technique will allow for robust statistical analysis and the determination as to whether particular light chains, and

moreover, particular CDR-L3s, pair preferentially with particular heavy chains --- either the particular IGHV segment or even a particular class of CDR-H3. A specific analysis of the primary antibody repertoire of IgM^+ naïve B cells in peripheral blood is predicted to reveal identical, public CDR-L3s, and even identical full-length IGL chains, but paired with multiple distinct IGH chains; thus, indicating non-clonally related sequences which have arisen independently in the primary repertoire. Imminent evidence that such must be the case is the very fact that the IGH repertoire is documented to be extensively diversified [18], [71] and, therefore, the restriction of IGL diversity we observe here can only be accommodated if it pairs promiscuously with multiple IGH chains and their CDR-H3s. Such a striking observation would immediately imply an unexpected IGL-mediated mode of selection, independent of the IGH chain, which recruits IgM^+ naïve B cells into the pool of the primary antibody repertoire.

Chapter 4: Systematic Characterization and Comparative Analysis of the Rabbit Immunoglobulin Repertoire

Specifically for the work described in this chapter, K.H. Hoi prepared and supplied the reagents, materials, bioinformatics scripts and analytical tools. And, K.H. Hoi performed the experiments together with J.J. Lavinder, S.T. Reddy, and Y. Wine.

This chapter is reproduced with minor modifications from its initial publication:
J. J. Lavinder, K. H. Hoi, S. T. Reddy, Y. Wine, and G. Georgiou, "Systematic Characterization and Comparative Analysis of the Rabbit Immunoglobulin Repertoire," PLoS ONE, vol. 9, no. 6, p. e101322, Jun. 2014.

Manuscript author contributions:
J.J. Lavinder and G. Georgiou initiated the study and designed experiments; K.H. Hoi supplied the reagents, materials, scripts, and analytical tools; J.J. Lavinder, K.H. Hoi, S.T. Reddy, and Y. Wine performed the experiments; J.J. Lavinder and G. Georgiou interpreted the data; J.J. Lavinder, K.H. Hoi, and G. Georgiou wrote the manuscript.

Acknowledgements:
The authors would like to acknowledge Dr. Scott Hunicke-Smith for assistance with NGS, Constantine Chrysostomou for assistance in data analysis, Bob Glass for assistance with rabbit immunization and bone marrow isolation, Professor Gregory Ippolito for reading the manuscript, and Prof. Andrew Ellington and Brent Iverson for useful discussions and comments.

INTRODUCTION

B cell development and repertoire diversification vary significantly among vertebrate species [160]. Diversification of the Ig repertoire occurs through the combinatorial joining of numerous V, D, and J gene segments for the Ig heavy chain (or just V and J gene segments in the case of Ig light chains) through several mechanisms collectively referred to as VDJ recombination, followed by somatic mutagenesis upon subsequent B-cell encounter with foreign antigen. Compared to humans and mice, which use a diverse assortment of germline VH gene segments during VDJ recombination of the

heavy chain, the rabbit IgH repertoire displays highly restricted VH gene segment usage. Earlier studies had indicated that the majority of B cells in the rabbit utilize the VH1 gene, the most D-proximal VH locus [161]. VH1 Igs are serotypically VHa-positive, and there are three distinct VHa allotypic lineages (a1, a2, and a3) [162], [163]. In addition, approximately 10-20% of expressed Ig in rabbits are serotypically VHa-negative (VHn) [163], [164]. The VHn Ig genes that have been annotated in rabbits (VHx, VHy, and VHz) are encoded by loci significantly upstream (>100 kb) of the VH1 gene locus [165]. Recently, sequencing of the rabbit genome has enabled the identification of germline Ig elements in a Thorbecke inbred rabbit [166]. Overall, >300 VH-like gene sequences were identified within 79 unplaced genomic scaffolds (i.e. unknown chromosomal locations). The large number of previously unannotated VH-like sequences identified within the a1/a2 Thorbecke rabbit, as well as previously identified sequences from latent heavy chain allotypes [161], [167], clearly demonstrate the complexity of the germline Ig repertoire. However, because the sequenced Thorbecke rabbit was heterozygous at the IgH locus (a1/a2 based on mapping of the VH1 gene), the actual number of distinct VH gene elements in the haploid genome is unclear.

Another major source of Ig repertoire diversity derives from the somatic introduction of non-templated nucleotides into the imprecise junctions formed by the variable ligation of recombining V-D and D-J gene segments—a process known as N-nucleotide addition. This hypervariable V-*N*-D-*N*-J interval defines CDR3 of the heavy chain (CDRH3). Species such as cattle have extremely long CDRH3s [168] as a result of increased levels of N-nucleotide addition. Longer CDRH3s not only create a more expansive and diverse sequence space in the Ig repertoire, but may also hold unique functional relevance in protection against disease [169]. For most mammalian species, N-

nucleotide addition during VJ recombination of the light chain is limited and therefore junctional diversity in the light chain is much less pronounced compared to the heavy chain; however, rabbits have been shown to have light chain CDR3s (CDRL3s) that are unusually longer and more diverse, indicating significant N-nucleotide addition during light chain VJ recombination [170].

After VDJ recombination, the naïve Ig repertoire in rabbits is further diversified in the first 2 months of age by extensive somatic mutagenesis in the gut-associated lymphoid tissue (GALT) [171], through both somatic hypermutation (SHM) and gene conversion events [172], both of which have been shown to be dependent upon the exposure of the naïve B cell repertoire to the gut microflora [173]. Ig gene conversion is employed not only by rabbits, but also by other species including chickens and involves the non-reciprocal homologous recombination of upstream donor V gene loci into the recombined VDJ (and VJ) locus. Like SHM, Ig gene conversion is mediated through the enzyme *a*ctivation *i*nduced cytidine *d*eaminase (AID) [174] and thus is often found to occur proximal to hotspot AID motifs conserved within germline V genes. In chickens, gene conversion has been shown to be the dominant mechanism of AID-mediated mutagenesis [175] and involves a single functional VH and VL gene undergoing gene conversion with numerous upstream VH and VL pseudogenes, respectively [176]. In rabbits, however, the upstream loci are a mix of functional V genes and pseudogenes that can serve as potential donor sequences in gene conversion events. The fundamental properties of gene conversion events and the relative extent to which gene conversion plays a role in rabbit Ig diversification is not entirely clear, mostly due to limitations in sampling and difficulty in precise, automated identification of gene conversion events in highly mutated Ig sequences.

Here, we present a thorough characterization of the expressed rabbit IgG repertoire. We identify several unannotated functional rabbit germline VH and VL germline gene sequences and provide a comprehensive survey of the salient features of the rabbit Ig repertoire. We estimate the gene conversion frequency in the rabbit and demonstrate that it is significantly less than that observed in the chicken repertoire and, not surprisingly, much greater than that observed in humans and mice.

MATERIALS AND METHODS

Ethics Statement

Three New Zealand white (NZW) rabbits and one white leghorn chicken were used for this work, as approved through the Institutional Animal Care and Use Committee (IACUC) of the University of Texas at Austin (protocol AUP-2011-00016). All efforts were made to ensure animal welfare and minimize suffering in accordance with the United States Department of Agriculture (USDA) Animal and Plant Health Inspection Service (APHIS) Guidelines for animal care and husbandry.

Isolation of B cells from immunized rabbits, chicken, mouse, and human

At sacrifice, rabbit femoral bone marrow (BM) cells were isolated and approximately 100 ml blood was collected into heparin tubes. Blood aliquots of 20 ml were gently layered over 20 ml of Histopaque 1077 (Sigma, MO, USA) and centrifuged in a swinging bucket rotor at 400*g*, 45 min at 25°C (Beckman Coulter). The serum was removed from the top of the gradient and stored at -20° C. PBMCs were isolated from the intermediate layer. Each collected tissue (BM and PBMC) was processed as previously

described [83], with the exception that the PBMCs did not require red blood cell lysis after gradient centrifugation. $CD138^+$ cells were isolated as previously described [84]. PBMCs or $CD138^+$ BM plasma cells (PCs) were centrifuged at 930x*g*, 5 min at 4°C. Cells were then lysed with TRI reagent (Ambion, TX, USA) and total RNA was isolated according to the manufacturer's protocol in the Ribopure RNA isolation kit (Ambion). RNA concentrations were measured with an ND-1000 spectrophotometer (Nanodrop, DE, USA).

For the chicken, total RNA was prepared from splenic tissue of a white leghorn chicken using TRIzol reagent (Life technologies) and purified with RNeasy Micro Kit (Qiagen, CA). cDNA was generated from total RNA using oligo(dt) according to the manufacturer's protocol (Superscript II First strand Synthesis kit, Life Technologies), PCR-amplified as described previously [177] using chicken IgY-specific primers listed in Table 7, and sequenced using the 2 × 250 paired end MiSeq Next Generation Sequencing (NGS) platform (Illumina, San Diego, CA, USA). The two IIlumina 2x250 output files were aligned using FLASH [98] and CDRH3 and full-length VH sequences were determined using in-house probabilistic model [83] for delimiting the CDRH3 regions based on *Gallus gallus* Ig sequences found in NCBI Genbank.

Amplification and high-throughput sequencing of rabbit VH and VL gene repertoires

Approximately 0.5 μg of ethanol precipitated RNA was used for first-strand cDNA synthesis according to the manufacturer's protocol for 5' RACE using the SMARTer RACE cDNA Amplification kit (Clontech, CA, USA). The cDNA reaction was diluted into 100 μl of Tris-EDTA buffer and stored at -20°C. 5' RACE PCR

amplification was performed on the first strand cDNA to amplify the VH repertoire with the kit-provided, 5' primer mix and 3' rabbit IgG-specific primers RIGHC1 and RIGHC2 (Table 7). The rabbit VL repertoire was amplified via 5' RACE, using a 3' primer mix specific for both the Vκ and Vλ rabbit constant regions. The VL primers comprised 90% RIGκC mix and 10% RIGλC mix (Table 7) to approximate known ratios of light chain isotypes in rabbits. Reactions were carried out in a 50 μl volume by mixing 35.25 μl H_2O, 5 μl 10X Advantage-2 PCR buffer (Clontech), 5 μl 10X Universal Primer A mix (Clontech), 0.75 μl Advantage-2 polymerase mix (Clontech), 2 μl cDNA, 200 nM V_H or V_L primer mix, and 200 μM dNTP mix. PCR conditions were: 95 °C for 5 min, followed by 30 cycles of amplification (95 °C for 30 sec, 60 °C for 30 sec, 72 °C for 2 min), and a final 72 °C extension for 7 min. The PCR products were gel-purified to isolate the amplified VH or VL DNA (~500 bp). 100 ng of each 5' RACE amplified VH or VL DNA was processed for Roche GS-FLX 454 DNA sequencing according to the manufacturer's protocol.

Primer Name	Sequence	Description of use
RIGHC1	CAGTGGGAAGACTGACGGAGCCTTAG	Rabbit IgG CH1 reverse V_H primer mix (equimolar)
RIGHC2	CAGTGGGAAGACTGATGGAGCCTTAG	Rabbit IgG CH1 reverse V_H primer mix (equimolar)
RIGκC1	TGGTGGGAAGAKGAGGACAGTAGG	Rabbit Igκ reverse primer mix (90% of mix)
RIGκC2	TGGTGGGAAGAKGAGGACACTAGG	Rabbit Igκ reverse primer mix (5% of mix)
RIGκC3	TGGTGGGAAGAKGAGGACAGAAGG	Rabbit Igκ reverse primer mix (5% of mix)
RIGλC1	CAAGGGGGCGACCACAGGCTGAC	Rabbit Igλ reverse primer mix (equimolar)
RIGλC2	GTGAAGGAGTGACTACGGGTTGACC	Rabbit Igλ reverse primer mix (equimolar)
RIGλC3	GAGGGGGTCACCGCGGGCTGAC	Rabbit Igλ reverse primer mix (equimolar)
Chicken VH1	GCCGTGACGTTGGACGAGTCC	Chicken VH1 forward primer
Chicken IgY	GGAGGAGACGATGACTTCGGTCCC	Chicken IgY reverse primer

Table 7: Primers used to amplify IgH and IgK/Igλ repertoire

All 454 data were first processed using the sequence quality and signal filters of the 454 Roche pipeline and then subjected to bioinformatics analysis that relied on homologies to conserved framework regions using IMGT/HighV-Quest Tool [68]. Additional filters were applied for full repertoire database construction as follows: (i) Length cutoff*:* full-length sequences were filtered by aligned amino acid lengths > 70 residues and aligned framework 4 region lengths > 2 residues; (ii) Stop codons: aligned amino acid sequences containing stop codons were removed.

IgBLAST alignment, Multidimensional scaling (MDS), and k-means analysis

An IgBLAST database for germline annotation of the rabbit IgG sequences was constructed using the following sequences: the IMGT rabbit V germline reference set that includes the allotypic a2 sequences in BAC clones AY386694 and AY386697 [178], allotypic a2 sequences from an Alicia rabbit (AF176997 through AF177016) [179], potentially latent IGHV (M12180, M60121, M60336) [167], [180], [181], allotypic a1 sequences VH1-a1 (M93171), VH3-a1 (M93177), and VH4-a1 (M93181) [182], and the allotypic a3 sequences VH1-a3 through VH7-a3 (M93173, M93176, M93179, M93183, M93184, M93185, M93186) [172], [182]. In addition to the IMGT rabbit reference set, initial IgBLAST database included VH8-a3 through VH11-a3 (L27311, L27312, L27313, L27314) [183], VHx (L03846) [184], and VHy (L03890) [184]. For light chain, the IMGT database was used without addition. IgBLAST alignments against the database were analyzed by bit score (and equivalently the number of called nucleotide mutations per sequence). Aligned (annotated to a certain germline) sequences with greater than 30 called mutations were extracted from this initial IgBLAST alignment and these poorly

aligned sequences were aligned using MUSCLE [185] multiple sequence alignment (BLOSUM80 substitution matrix, gap open penalty -15, gap extend penalty -3). For calculating distance matrices and performing MDS, the package bios2mds [186] in the R environment was used. The MUSCLE alignment was imported into R and the pairwise distance matrix calculation using the 'mat.dif' function, which computes a distance matrix based on pairwise differences between each sequence was performed. Metric MDS analysis of the pairwise distance matrix was performed using the function 'mmds', which reduces the dimensionality of the distance matrix into Euclidean space. These Euclidean values are analyzed by k-means silhouette scoring (function 'sil.score) and k-mean clustering (function 'Kmeans') to identify distinct sets of sequences that each derived from an unannotated germline Ig sequence. The sequences from each cluster are extracted and aligned in MUSCLE. For each derived cluster alignment, the consensus sequence was searched by BLASTn against the non-redundant nucleotide collection and the rabbit genome.

IMGT and IgBLAST repertoire analyses

Germline V gene assignments were derived from IgBLAST alignments against the database described above. Germline J gene assignments and CDR3 sequences (rabbit, mouse, and human) were derived from IMGT HighV-Quest alignments. Chicken CDR3 sequences were derived from a position weight matrix motif search of the FR3 and J region in chickens.

Gene conversion analysis

For rabbits, IgBLAST alignments of the NGS data sets was performed using custom BLAST databases for rabbit, as detailed above. For the chicken, the IgBLAST database included the functional VH1 sequence, along with 18 known VH pseudogenes [176]. For mouse and human, the IgBLAST-provided database was used. IgBLAST was used to assign the best-scoring germline VH reference sequence for each query sequence. To detect gene conversion events in the query, the assigned germline reference sequence was then scored against all other germline reference sequences in the IgBLAST alignment as follows (an example is shown in Figure 19): 1) For each VH germline in the alignment (each a possible donor VH sequence) except the assigned one, we used a scoring function that assigns a '1' at each position only if the putative donor VH matches and the assigned reference VH germline mismatches, a '0' at each position that both references either match or both mismatch, and a '-1' at each position that the assigned reference VH matches and the putative donor VH mismatches. 2) Search each scored putative donor VH for stretches of positions that score as '1', with a putative gene conversion event called only if three positions scoring '1' are uninterrupted by positions scoring '-1'. The gene conversion event boundaries were defined by positions scoring '-1' (long tract boundary) or by the most distal positions of the tract that score '1' (short tract boundary). Adjacent long tracts from the same donor VH are automatically combined by allowing long tracts with a shared boundary to connect. Positions of the alignment that have gaps in the query are scored as '0' in all putative donor VH scored positions. To exclude PCR crossover products or gene replacement events (single crossover events), all gene conversion events that start within the first 15 positions or end with the last 15 positions of the aligned VH gene are excluded (e.g. the gene conversion must be an internal double crossover event with sufficient sequence from the assigned

VH on each side). The donor VH selected represents the germline VH with the highest scoring tract (sum of the tract positional scores). P-values for the gene conversion events are scored as described [187], with the exception that all polymorphic sites are permuted during the permutation test. The p-values described here are local p-values calculated via 1000 iterations of positional permutation of the assigned and donor VH germlines. Only gene conversion events with a p-value below 0.05 (95% confidence interval) and a minimum tract score > 4 (to avoid effects of high SHM) are considered as high confidence events.

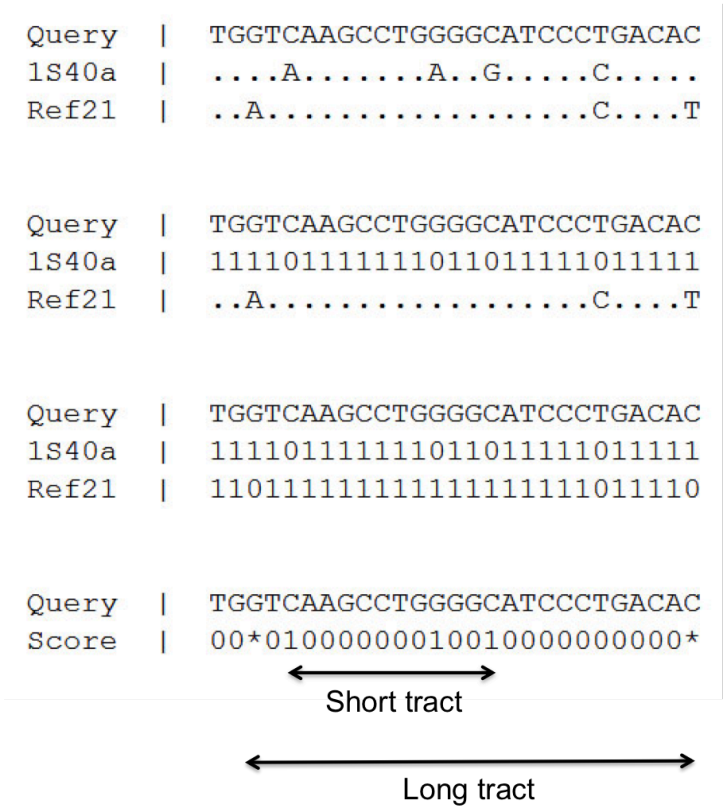

Figure 19: An example summarizing the scoring system.

Note: '*' denotes a score of -1

RESULTS

Identification of putative rabbit VH germline elements using multidimensional scaling of high throughput sequencing data

Total RNA was isolated from BM PCs and total PBMCs of three adult NZW rabbits. IgG heavy chain and Igκ/Igλ light chain cDNAs were amplified by 5' RACE using primers that annealed respectively to the CH1 or CK/Cλ constant region directly 3' of the J segment (detailed procedures in materials and methods section), and the resulting amplicons were sequenced by Roche 454 sequencing. 172,126 high quality reads corresponding to 88,830 unique heavy chain sequences across the three rabbits were obtained (Table 8).

Sample	reads	unique VH/VL amino acid sequences	unique CDRH3/CDRL3
Rabbit rab1 PBMC VH	16102	9447	5525
Rabbit rab1 Bone marrow PC VH	31136	19044	5954
Rabbit rab2 PBMC VH	24251	13459	7220
Rabbit rab2 Bone marrow PC VH	76510	34762	11564
Rabbit rab3 Bone marrow PC VH	24127	12118	5958
Rabbit rab1 Bone marrow PC VL	24489	10446	5629
Rabbit rab2 Bone marrow PC VL	17155	7487	4465
Rabbit rab3 Bone marrow PC VL	23761	12581	7139

Table 8: Summary of sequencing reads from 454 DNA sequencing

Germline VH usage was determined with IgBLAST [92] alignments using a custom database that included NZW rabbit germline sequences compiled from a number of sources [167], [172], [178]–[184] (see Materials and Methods). For the VHa sequences in all three rabbits, >99% were of the a3 allotype, strongly indicating that the cohort of NZW rabbits examined here is homozygous a3/a3 at the IgH locus. However, the

IgBLAST alignments revealed a non-normal distribution of VH germline alignment scores (Figure 20).

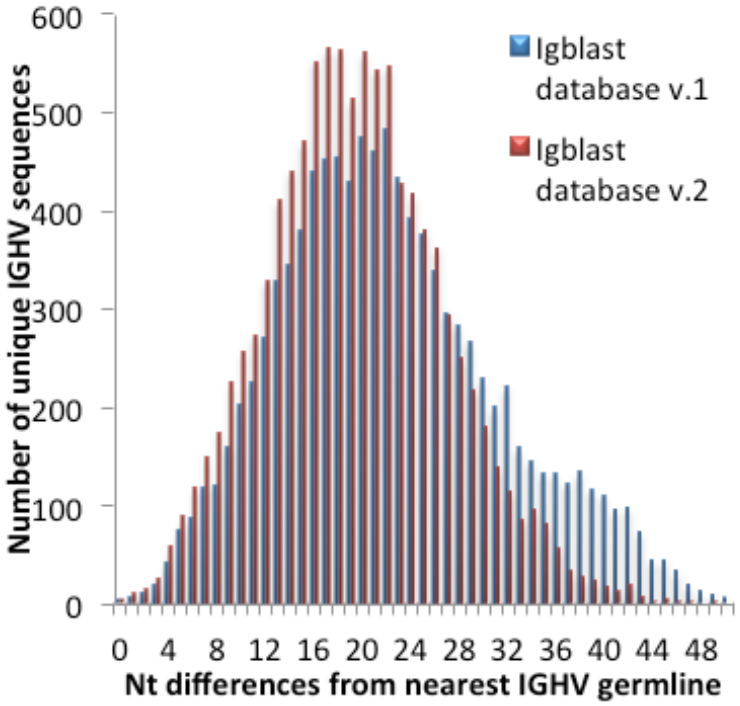

Figure 20: Comparison of IgBlast alignment before and after the addition of the putative sequences identified via MDS and k-means clustering

Note: before is v.1 in figure while after is v.2 in figure

Based on an analysis by Gertz et al. [166] revealing a number of unannotated germline elements in an a1/a2 Thorbecke rabbit, we hypothesized that the NZW rabbit germline database may be incomplete and thus lack the germline V gene sequences for these poorly scoring Ig alignments. MDS [188], a space-based method that has been used to identify patterns in distance matrices derived from multiple sequence alignments (MSAs) of large biological sequence data sets [186], [189], [190], was employed to

deduce putative germline V gene segments. MDS allows MSA distance matrices to be analyzed in Euclidean space, facilitating k-means clustering [191] of the sequences. In the case of somatically mutated Ig V gene sequences, the consensus sequence of each of these k-means defined clusters represents a putative germline V gene sequence. Figure 21 shows the MDS and k-means clustering of the poorly aligned VH gene sequences (higher than 30 nt differences from the nearest VH germline) in the NZW CCH1 immunized rabbit (CCH1 BMPC) repertoire as an example. For each of the other rabbits, the same pattern of four distinct VH clusters was identified. Each cluster of VH sequences was extracted and aligned, and the consensus sequence for each of the four clusters was compared across the three rabbits. Each of the four VH consensus sequences matched identically across all three rabbits, strongly supporting our hypothesis that the poorly aligned sequences are derived from unannotated germline VH elements encoded in the NZW rabbit genome.

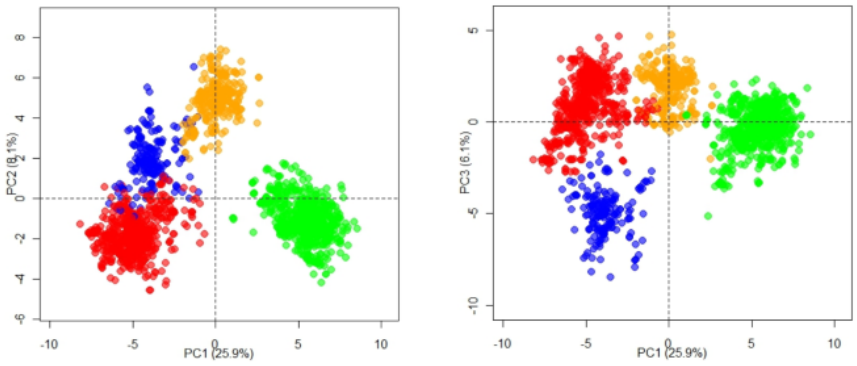

Figure 21: MDS and k-means clustering of low scoring alignments for CCH1 rabbit

Note: BMPC VH sequences: the left panel is PC1 vs PC2 and the right panel is PC1 vs PC3 while each colored groups is determined by k-means clustering of the Euclidean MDS-derived values

The four putative germline sequences identified by MDS and k-means clustering were searched by BLASTn to identify homology to publicly available rabbit genomic and transcript sequences (Table 9).

	accession number	mismatch/ total (nt)	Source
VHn3	M12386, NW_003160066	3/288, 4/288	(1) liver genomic DNA putative VH, (2) Thorbecke rabbit unplaced genomic scaffold
VHs1	JN131896, AY676808, NW_003161050	2/285, 1/285, 3/285	(1) imm-B cell genomic VDJ, (2) PBC mRNA LigApx rabbit, (3) Thorbecke rabbit unplaced genomic scaffold
VHx2	AF264452, AF264440, NW_003159519	0/288, 0/288, 3/288	(1,2) PBC mRNA LigApx rabbit, (3) Thorbecke rabbit unplaced genomic scaffold
VHn2	AF245499	10/288	mRNA from rabbit bone marrow and spleen

Table 9: Blastn results of the four putative VH germline sequences identified by MS and k-means clustering

For three of the four putative VH germline sequences, NZW rabbit genomic or transcript sequence matches were found that were identical or within 1-3 nucleotide differences. The closely matching transcript sequences (AY676808, AF264452, and AF264440) were derived from rabbits that have a ligated appendix (LigApx) [173], [192], which effectively eliminates SHM and gene conversion. Three of the four putative germline sequences contained a ^{70}WVN^{72} motif, consistent with VHa-negative (VHn) immunoglobulins (VHa sequences have a ^{70}WAK^{72} motif), while one sequence (VHs1) had a ^{70}SVK^{72} motif, which is predominant in VHs immunoglobulins (which are also VHa-negative) and ancestral to hares [193]. VHx2 was highly identical (281/288 nt) to the VHx32 allele previously annotated [184] and may represent a distinct VHx allele

(hence its designated ID). These four new putative germline sequences in the NZW rabbit were added to our existing NZW rabbit germline database (see Materials and Methods for full description) and using this updated database, IgBLAST was used to assign VH and JH germline usage (Figure 22). Consistent with earlier observations [84], [161], the VH1 gene is heavily utilized in all three rabbits, as is the VH4 gene, which is >97% identical to VH1. The VHa-negative sequences (combined) account for 12%, 22%, and 11% of the total IgG sequences in CCH1 BMPC, CCH1 PBMC and CCH2 BMPC rabbit respectively. All three rabbits also exhibit highly restricted JH usage, with JH4 accounting for 60-70% of the IgG repertoire.

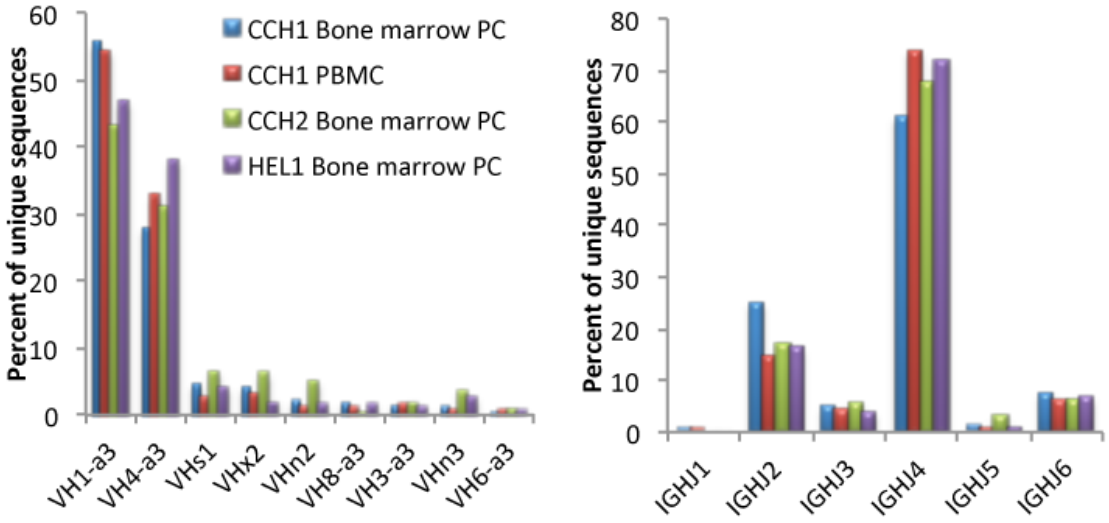

Figure 22: Heavy chain germline gene usage

Note: the left panel is the VH usage and the right panel is the JH usage

Vκ and Jκ usage in the rabbit

Similar to mice, rabbits utilize the kappa light chain isotype at a much higher frequency than the lambda isotype [194]. We amplified the light chain repertoire from BM PCs in all three rabbits using 5' RACE and sequenced the VL region using Cκ and Cλ specific primers. A total of 65,405 high quality reads and 30,514 unique sequences across the three rabbits were obtained (Table 8). As expected, the utilization of lambda light chain sequences sets was very low (<1%). Rabbit immunoglobulin kappa light chains have four allotypes: b4, b5, b6, and b9 [195]. For each of the three rabbits examined here, more than 98% of the unique VL sequences were of the b4 allotype, indicating this cohort of NZW rabbits was b4/b4 homozygous. Similar to the results of the VH IgBLAST alignments, Vκ gene alignment scores also revealed a non-normal distribution, with a group of sequences exhibiting significantly lower alignment scores as compared to the bulk of the Vκ sequences (data not shown). These poorly aligned sequences were examined more closely by MDS and k-means clustering as described above and in the Materials and Methods, and four new Vκ clusters were identified (data not shown). Two of the four putative Vκ germline sequences (NZWk57r and NZWk155g) were utilized in all three rabbits. NZWk57r and NZWk155g has also been detected in non-functional light chain sequences (VJ junction out-of-frame) in the bone marrow of a 1 day old b5/b5 NZW rabbit (i.e. early development when naive, unmutated Ig sequences are common in the rabbit) [170]. For the other two putative Vκ germlines, one was identified only in the CCH1 PBMC and CCH2 BMPC rabbit samples (NZWk807y), while the other was identified only in the CCH1 PBMC rabbit sample (NZWk529g). Nonetheless, all four cluster consensus sequences were also found by BLASTn analysis as either exact matches or differing by only 1 nt (NZWk807y) from previously identified germline genes in the Thorbecke inbred rabbit.

The four putative Vκ sequences were added to our existing NZW IgBLAST database, which was then used to assign germline Vκ usage (Figure 23). Contrary to the sharp germline restriction seen in the VH gene repertoire, Vκ gene usage is very diverse, with the top germline gene segment used at ~10-20% and 30 Vκ germlines utilized at least >1% (of total unique Vκ sequences) across the three rabbits. Jκ germline usage, on the other hand, is mostly restricted to the IGHK1_2 gene (~90%) and to a very small extent IGHK1_1 and IGHK2_2.

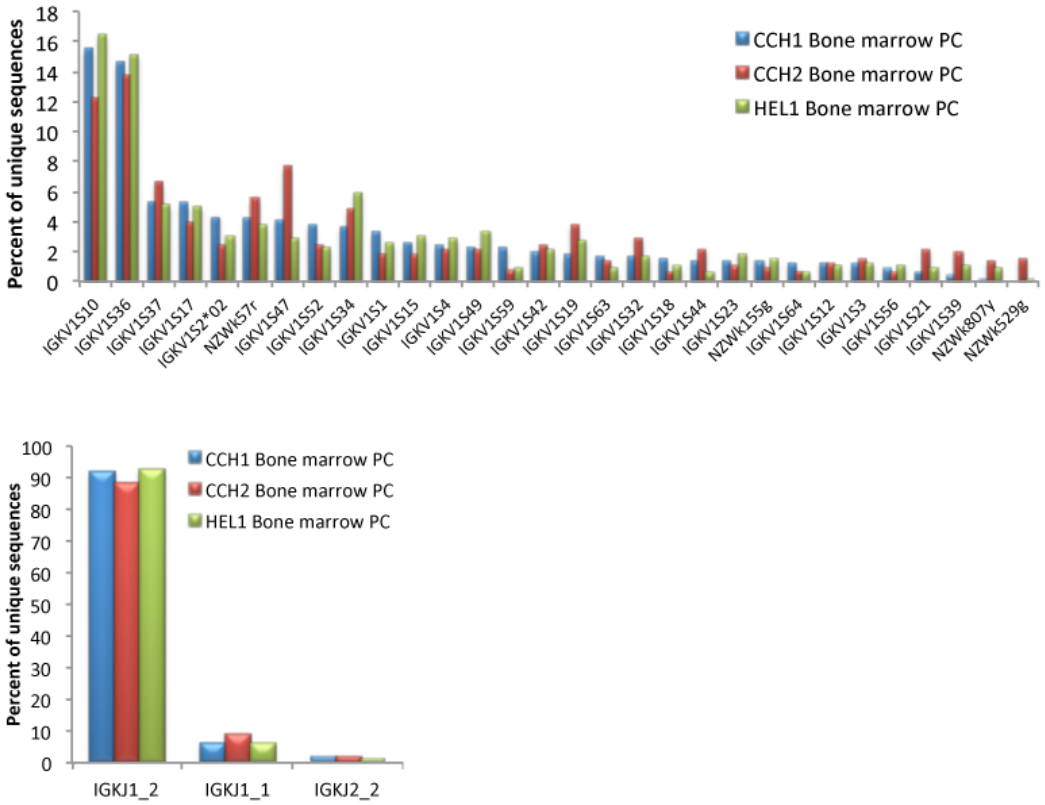

Figure 23: Light chain germline gene usage

Note: upper panel is the VK gene usage and the lower panel is the JK gene usage

Characterization of the CDRH3 and CDRL3 in the rabbit IgG repertoire as compared to other species

In addition to the rabbit NGS data set, we also analyzed human [85], mouse [83], and chicken NGS data sets to compare and contrast repertoire characteristics across species. For the chicken, we obtained 320,468 high quality VH sequence reads (231,165 unique VH amino acid sequences) from the splenic B cell repertoire of a white leghorn chicken using the Illumina MiSeq 2x250 NGS platform. A comparison of the CDRH3 length distribution is shown in Figure 24. Rabbit IgG CDRH3 lengths are intermediate (mean = 14.8±3.6 aa, mode = 13 aa) relative to mice (mean = 11.1±2.0 aa, mode = 10 aa), humans (mean = 15.3±4.0 aa, mode = 15 aa), and chickens (mean = 17.9±2.8 aa, mode = 16 aa). The length distribution of the CDRH3 for all unique IgG sequences was similar across all three rabbits (Figure S3). For CDRL3, mice and humans both exhibit very little junctional diversity and are severely restricted in length, with the vast majority of CDRL3s for both species being 9 ±1 amino acids (Figure 24); However, the rabbit exhibits significant junctional diversity in the CDRL3, with a wide distribution of CDRL3 lengths (range: 5aa – 16aa) and a much greater mean length, equal to 12 ±1.6 aa. The amino acid composition of the rabbit Ig CDRH3 is dominated by tyrosine (Y), glycine (G), and aspartate (D) which together represent half (49%) of the amino acid usage in the CDRH3 loop (Figure 24), while the top five amino acids used (GYDAS) represent a full two-thirds (66%) of the amino acid usage. In that regard, the overall amino acid utilization in the rabbit is highly similar to the other species, consistent with earlier observations [196] that the average hydrophobicity of CDRH3—and, hence, the center of the antigen binding site—is conserved across evolution to be slightly hydrophilic and enriched for glycine, serine and tyrosine.

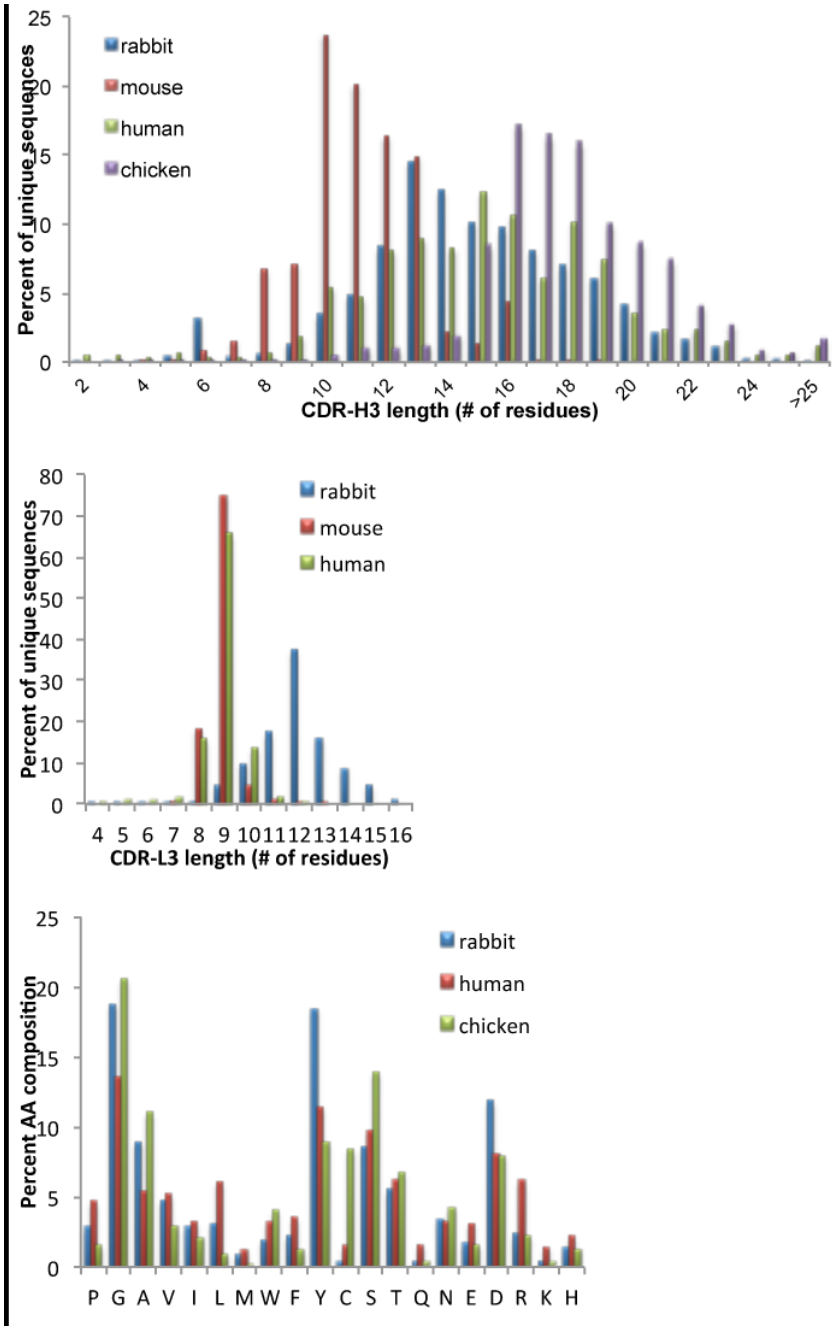

Figure 24: Cross species difference in the characterization of the CDRH3 and CDRL3

Note: the top panel is CDRH3 lengths, middle panel is the CDRL3 lengths, and the bottom panel is amino acids composition

Nevertheless, when compared to other species, the CDRH3 amino acid composition in rabbits does show some distinct features. Human CDRH3s use glycine and tyrosine at a much lower frequency than that seen in rabbits. Chicken CDRH3s have less tyrosine (~ 2-fold less than rabbits) but utilize much higher cysteine content (~ 5–10-fold higher than humans or rabbits). The higher utilization of cysteine residues in the chicken CDRH3 repertoire has previously been shown to be important for stabilizing (by disulfide bonds) the longer CDRH3 loops seen in chickens [197].

Diversification of the rabbit IgG repertoire by SHM and gene conversion

The rabbit Ig repertoire is known to undergo extensive AID-mediated mutagenesis (via both SHM and gene conversion) early on in development when the antigen-inexperienced naïve B cell repertoire migrates from the bone marrow to the GALT [165]. Earlier studies with rabbits lacking an established gut microflora demonstrated significantly reduced levels of AID-mediated diversification of the repertoire, with most Ig having sequences that approximate the germline elements from which they are derived [173], [192].

We compared the overall level of mutation (combined SHM and gene conversion) within the IgG repertoires of rabbits, chicken, mice and humans (Figure 25). The mutational load varied as follows: chicken>rabbit ≈ human>mouse. It should be noted that the reported mutational load is a combination of both biological processes mediated by AID and inherent PCR/sequencing error, which has been reported to be approximately 1% for both 454 GS-FLX [198] and Illumina MiSeq sequencing [77].

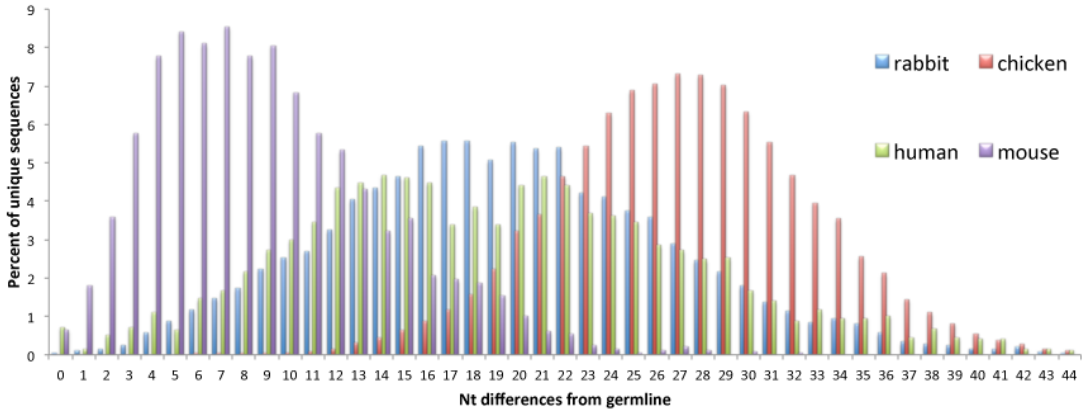

Figure 25: Comparison of overall nucleotide deviations in the VH sequences across species

To determine the relative contribution of gene conversion to the diversification of the primary repertoire, we developed a script that searches Ig sequences for tracts of putative gene conversion events (see Appendix). Gene conversion tracts are detected as a contiguous block of nucleotides within a query Ig sequence that closely matches a different germline element (e.g. not the query's assigned germline element) in the IgBLAST database. Additionally, to rule out possible PCR template switching artifacts, the gene conversion tracts were required to be bound on each end by positions (tracts) that match the query's assigned VH germline sequence (i.e. the gene conversion event was not contiguous with the 5' or 3' ends of the sequence). Additionally, minimum scoring and p-value thresholds were applied as described in the methods. Strict statistical thresholds were set to ensure that the identified gene conversion events were highly significant and not attributed to high loads of point mutation. For these reasons, the

reported frequencies of gene conversion events should be considered as a lower bound of the actual biological frequencies (Table 10).

	unique seqs	% with GC	max GC score	avg tract (nt)
CCH1 rabbit BMPC IgG VH1	10680	23	17	59
Chicken Spleen IgY VH1	10000	70	23	79
Mouse BMPC IgG VH1	946	0.1	5	6
Human PBMC IgG VH3	1028	2.5	7	39

Table 10: Gene conversion comparative analysis across species

The vast majority of unique chicken IgY sequences examined (70%) display evidence of gene conversion events. In rabbits, 23% of IgG sequences were the products of gene conversion. There have been previous, although somewhat controversial, indications suggesting gene conversion occurs in humans and mice as well, albeit at a much lower frequency [199]–[201]. We find that, in the mouse, putative gene conversion events are nearly absent, with an estimated frequency of 0.1% of all unique IgG sequences. Whereas an earlier analysis of gene conversion in a small set of human IgG sequences indicated that ~7% (8 out of 121) display evidence of having undergone gene conversion [201], our present analysis of a much larger data set revealed a lower frequency of 2.5%. We note that, in humans and mice, the low p-values ($p<0.05$) in the detection of gene conversion events suggest that these are high confidence identifications despite the fact that the average tract lengths detected were significantly lower than those in the rabbit and chicken (Table 10).

The frequencies of donor germline VH usage for gene conversion in the rabbit are largely unknown. Figure 26 shows the donor germline VH usage for query sequences that were assigned by IgBLAST to one of three heavily utilized germline VH gene segments in the rabbit (VH1, VHs1, and VHn3). Because gene conversion occurs through homologous recombination, the frequency is heavily dependent on donor VH sequence homology and proximity. High homology donor VH genes directly upstream of the assigned VH reference (e.g. the VH germline originally used during VDJ recombination) are expected to be used in gene conversion more frequently than donor genes that are more distal or less homologous. The donor germline usage for VH1 is consistent with this expectation, with the genes directly upstream being used as donors for gene conversion more frequently than those more distal to VH1. The two VHa-negative sequences (VHs1 and VHn3) have very different patterns of germline VH donor usage. The genomic location and organization of these two VHa-negative elements are not known, but it is clear that VHs1 must be downstream of VHn3 as it heavily utilizes VHn3 as a donor sequence for gene conversion.

The tract lengths and start/end residue numbers of the gene conversion events for assigned VH1 sequences are shown in Figure 26 middle panel and 26 bottom panel. The majority of gene conversion tracts in rabbit IgG are under 30 bp in length, although some identified tracts are much longer (>120 bp). As expected for AID-mediated events, the gene conversion tracts have start and end positions that mostly localize to the CDRH1 and CDRH2 regions of the V genes, where a number of conserved AID hotspot motifs are located. These CDRs, along with CDRH3, constitute a large amount of the paratope in antibodies and thus are strongly mutated and selected during the affinity maturation process.

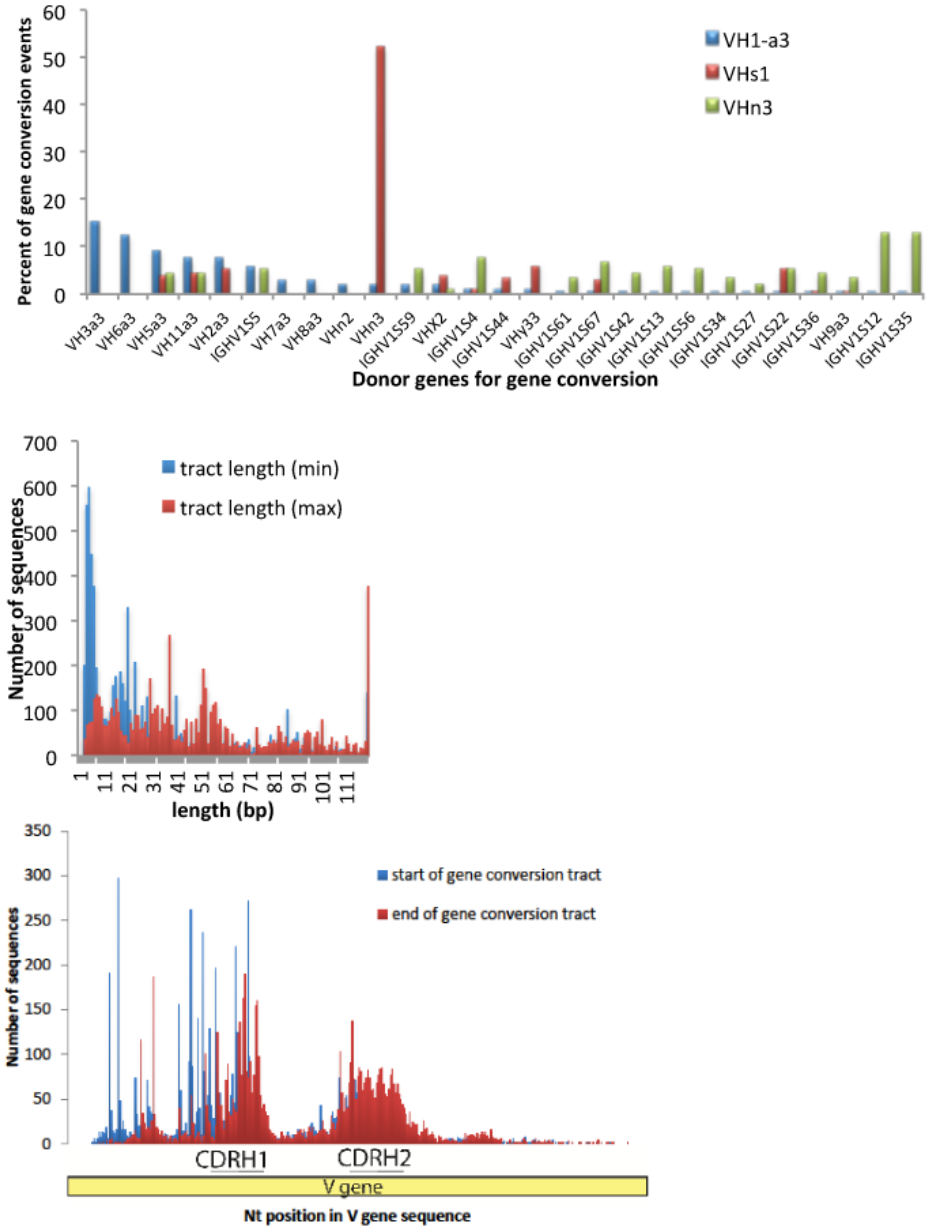

Figure 26: Gene conversion analysis

Note: top panel describes donor genes utilization in gene conversion, middle panel describes gene conversion tract length, and bottom panel describes the nucleotide positions along the VH gene sequence where the gene conversion recombination events start and stop with respect to short tract.

DISCUSSION

The vertebrate adaptive immune system is unparalleled in its ability to sample the depths of protein sequence space for the production of high-affinity antibodies endowed with exquisite specificity. Not only are antibodies extremely useful in the lab as affinity reagents, but they also represent the fastest growing sector of the biologics drug market, with annual global sales for monoclonal antibodies approaching $50 billion [202]. This has resulted in an increased interest for mining the antibody repertoires within vertebrates in a systematic, high resolution manner, something afforded by increasingly economical NGS technologies that enable the collection of thousands to millions of DNA sequences in a single sequencing run. Several species' Ig repertoires have been characterized by NGS to date [20], [69], [83], [130], [168], [203]. In this report, we used 5' RACE-amplification of rabbit IgG and Igκ/Igλ transcripts, followed by NGS and bioinformatics analyses, to elucidate key features of the repertoire. We provide evidence that the existing rabbit germline VH gene database, as annotated from a number of sources [167], [172], [178]–[184] (see Materials and Methods), is incomplete. This was not surprising based on previous estimations of the number of Ig germline elements in the rabbit and also a very recent survey of Ig germline elements detected in the genome of a Thorbecke inbred rabbit [166].

There are typically two types of approaches for examining sequence relationships in the multiple alignments of homologous sequences: (1) tree-based methods (e.g. phylogenetics) and (2) space-based methods that, unlike phylogenetics, do not infer a hierarchical or a specific structure within the sequence alignment. For the assignment of

germline sequences, space-based methods provide a statistical framework for comparing and clustering the sequences based on pairwise identities or similarities. MDS is a space-based method that allows the pairwise distances in the multiple sequence alignment to be reduced to a small number of principle components that aid in clustering the data within Euclidean space. This type of analysis applied to large Ig sequence data sets allows accurate genotyping of the germline elements within the species simply based upon the detection of highly frequent shared polymorphisms observed across individuals [142]. We show that MDS combined with k-means clustering provides an efficient approach towards discovery of new Ig germline elements in NGS data sets, even with repertoires that exhibit high loads of mutations, as is the case with the rabbit IgG repertoire where a large fraction of Ig sequences deviate significantly from the germline due to gene conversion events. MDS combined with k-means clustering could be successfully applied to a multitude of species for which the germline Ig loci are poorly annotated.

The large sample size provided by NGS also allows the diversification mechanism of Ig repertoires to be analyzed in great detail. We show that in the rabbit, the frequency of gene conversion is significantly lower than in the chicken. Consistent with this finding, it had been previously reported that chickens depend on gene conversion as the primary mechanism of Ig diversification and that SHM play a smaller role [204]. In rabbits, the chromosomal organization of VH gene elements is quite complex, with many VH germline genes located in genomic regions far removed from the commonly utilized VH1 germline gene. This may effectually limit the relative frequency of gene conversion, as gene conversion of VH1 is limited mostly to those donor genes directly upstream. Further, several of these upstream donor genes are functional, whereas in chickens there exists a single functional germline VH and a pool of upstream pseudogenes that are used

exclusively as donor genes for gene conversion. Interestingly, and consistent with earlier data [199], we report a detectable amount of gene conversion in the human IgG repertoire, but not in the mouse. The gene conversion tract lengths are significantly lower in the expressed human IgG repertoire as compared to the rabbit and chicken, but nonetheless are of high statistical confidence ($p<0.05$). This finding argues that gene conversion needs to be explicitly taken into account in the analysis of the antibody repertoire.

Chapter 5: Assessment of the circle sequencing technology in detecting true sequence variants

INTRODUCTION

The innate immunity, usually provoked in the early phase of pathogen exposure, is the "first responder" in host protection. However, this immediate response is usually short-term and non-specific in nature. Adaptive immunity, on the other hand, is usually long-term and specific against pathogen re-exposure albeit the need of an incubation period for protection development. Hence, adaptive immunity is a critical arm of immunity, providing effective long-term immune protection as well as immunological memory. Chapter 1 provides a detailed description of numerous mechanisms employed by adaptive immunity to ensure that antibody-producing cells (B cells) can accommodate specificity for a large diversity of potential pathogens. The diversity of the antibody repertoire or the B cell repertoire is reflective of the immune status; hence, measuring the antibody repertoire or the B cell repertoire enables the evaluation of the immunological landscape. For example, researchers can identify unique antibody repertoire signatures associated with the various dysfunctional or infectious diseases, such as chronic lymphocytic leukemia, rheumatoid arthritis, influenza, and dengue, etc. [69], [86], [101], [205]–[210]. Particularly, due to its uniqueness in amino acid sequence, the evaluation of the 3rd complementarity-determining region on the heavy chain (CDRH3) has shown to be most versatile for studying the immunological landscape; therefore, it can be utilized as a measure of the clonal cell population [211], [212]. Furthermore, CDRH3 sequence is referenced as a unique marker to deconvolute the serological antibody repertoire [84], [85]. Therefore, in order to accurately characterize the repertoire, it is essential to develop

methods which separate true sequence variation in the CDRH3 from sequence artifacts caused by sequencing errors.

Variation due to sequence artifacts can be introduced at multiple stages of library preparation: (1) error can be introduced in reverse transcription, (2) error can be introduced during PCR amplification, and (3) error is introduced during next generation sequencing. A commonly accepted mutation rate for the commercially available reverse transcriptase (MMLV-RT or Invitrogen SuperScript) is around 10^{-4} to 10^{-5} error per base [213], [214] and the DNA polymerase commonly used has similar range of error rates [215]. As for the sequencing platform, the rate and the type of error differ respectively. The Roche 454 pyrosequencing and Life Technologies Ion Torrent are both prone to insertion/deletion in the homopolymeric region while the Illumina HiSeq/MiSeq are prone to substitution [216], [217]. The error rates for each sequencing platform has been reported to be 1.4% for 454, 1.2% for Ion Torrent, and 3.2% for Illumina [217], [218]. Although the Illumina platform tends to incur higher error rates, its costs and sequencing depth have provided the most balanced option suitable for many studies.

One advantage of great sequencing depth is substantiated by improved sequencing reads fidelity through redundant confirmation [219]. Redundant reads can be especially effective at subsiding random mutation introduced during the sequencing process. The mutation reduction is achieved by negating the relatively few and scattered sequencing mutations from the consensus read where the correct bases are usually dominant. Due to the unavailability of a well-defined reference template in the CDR3 region of an antibody, an independent identifier must be used to trace redundant read products to its originating template sequence. Various groups have used barcoded primers to tag the

sequencing library or the first strand cDNA. This barcode can be used to group related-redundant reads identified in the bioinformatics process post sequencing [220]–[223]. To obtain a sufficient number of redundant reads per unique barcode, barcoding library templates requires a good balance between barcode primers, amount of input materials, and the number of amplification cycles. For example, in the study by Shugay et al. [223], only 2 cycles of amplification were performed prior to sequencing. After sequencing 2 million reads, the group was able to recover about 11,000 good unique barcodes. The barcoding method is direct but it might require individual optimization for each separate sequencing experiment and the cost-barrier might still be too rigid for wide-adoption. Alternatively, a method called circle sequencing [224], can generate tandem repeats of a template, and therefore, bypass the need to generate a sufficient library of independent sequences carrying the same barcodes. Moreover, even in the case of rare read instances, physically-linked tandem repeats can still ensure redundancy confirmation instead of being unaccounted for due to low redundancy recovery rate in the barcoded case. Due to the low barrier of sequencing depth requirement in obtaining redundant reads, circle sequencing appears to be a good methodology for detecting true variants in the CDRH3 sequences. Thus, this project herein will characterize circle sequencing comparing to the baseline conventional sequencing to identify differences in sensitivity for detecting true CDRH3 variant from the artifactual one.

MATERIALS AND METHODS

Blood sample collection

Blood samples from each healthy donor were collected no more than twice a month. For each venipuncture blood draw, less than 50 mL of blood was collected. The

blood collection protocol has been approved by the Institutional Review Board (IRB) at the University of Texas at Austin (protocol number 2012-08-0031). Additionally, sourced leukocytes pooled from anonymous healthy donors were acquired through the Gulf Coast Regional Blood Center (Houston, TX, USA).

Naïve B cells enrichment using magnetic activated cell sorting

Peripheral blood mononuclear cells (PBMC) were isolated from blood as described previously [18]. Briefly, whole blood was mixed in a 1:1 ratio with DPBS; 30mL of the mixture was then added slowly to a falcon tube containing 15 mL of Histopaque-1077 (Sigma-Aldrich, MO, USA) at room temperature. Separation by centrifugation at 800 xg was performed at room temperature for 15 minutes with the brake turned off. After centrifugation, whole blood was separated into four distinct layers. These layers from top to bottom are: Plasma, PBMC, Histopaque-1077, and pelleted Erythrocytes. The top plasma layer was decanted to allow convenient access to the PBMC layer and the PBMC layer was collected using Pasteur pipette. Freshly collected PBMCs were split and stored in 50 mL Falcon tubes, diluted to 50 mL using DPBS, and spun at 300 xg with brake for 10 minutes at room temperature. After centrifugation, the supernatant was discarded and pelleted PBMCs were resuspended with DPBS and combined. The combined PBMCs were then adjusted to a cell density of about $5x10^7$ cells/mL in 1 mL aliquots. For each aliquot, the cells were placed in a 5 mL polystyrene tube and 50 uL of EasySep Human Naïve B cell Enrichment Cocktail (STEMCELL TECHNOLOGIES INC, BC, Canada) was added. The cells/cocktail mixture was well-mixed and incubated at room temperature for 10 minutes. Then, 250 uL of well-suspended EasySep D Magnetic Particles was added to the cells/cocktail mixture

and incubated at room temperature for 5 minutes. Afterwards, the cell suspension was brought up to a total volume of 2.5 mL with DPBS and mixed well. The tube was then placed in the EasySep magnetic holder and set-aside for 5 minutes. In one continuous motion, the enriched naïve B cells were poured off into fresh tube leaving behind non-B cells labeled with magnetic particles. The enriched B cells were then added to TRI Reagent (Ambion, TX, USA) in a 1:2 ratio to make final volume to about 1 mL.

Total RNA purification

The enriched naïve B cell/TRI Reagent suspensions were used for total RNA purification with the RNeasy MinElute Cleanup Kit (Qiagen, MD, USA). The protocol provided in the kit was as follows. Briefly, 200 uL of chloroform was added to the 1 mL cell/TRI Reagent mixture and vortexed for 15 seconds. The samples were then set-aside at room temperature for 5 minutes. The samples were centrifuged at 12,000 xg at 4 ºC for 10 minutes. The supernatant layer of about 400uL was transferred to a fresh tube and one volume of 70% ethanol was added to the transferred supernatant and mixed thoroughly. The supernatant/ethanol mixture was then added to the RNeasy MinElute spin column placed in a 2 mL collection tube. After closing the tube gently, the column was spun at 8000 xg for 30 seconds and the flow-through was discarded. The column was washed with 700 uL of the RWI buffer and spun at 10,000 xg for 15 seconds and the flow-through was discarded. The column was then placed in a fresh 2 mL collection tube and 500 uL of RPE buffer was added. The column was then centrifuged at 10,000 xg for 15 seconds. After discarding the flow-through, 500 uL of 80% ethanol was added to the column. The column was subsequently spun for 2 minutes at 10,000 xg. After discarding the flow-through, the column was spun with the cap open at full speed for 5 minutes. The

column was then transferred to a 1.5 mL eppendorf collection tube and 25 uL of DI water was added to the column for elution. The column was allowed to incubate for about 1 minute before being spun at full speed for 1 minute for the total RNA elution. The eluted 25 uL total RNA was then ready for subsequent steps or for storage at -80 ºC freezer.

The generation of cDNA amplicon with reverse transcription and PCR amplification

The isolated total RNA was reverse transcribed using the Super Script III reverse transcriptase (Invitrogen, USA) to generate the first strand cDNA. An initial mixture was prepared with following components: total RNA template at a concentration of approximately 250 ng per 10 uL solution, 1 uM of IgM reverse primer (5'-AACGGGGAATTCTCACAGGAGAC-3') in 2 uL, 10 mM dNTP in 1uL, and DI water to make up final volume to 13 uL. The mixture was heated to 65 ºC for 5 minutes and subsequently cooled on ice for at least 1 minute. The following components were added while the mixture was still on ice: 5x First-strand buffer in 4 uL, 0.1M DTT in 1 uL, RNaseOUT (Invitrogen, USA) in 1 uL, and Invitrogen Super Script III (Invitrogen, USA) in 1 uL. The final mixture in 20 uL total volume was then subjected to thermocycling with the following temperature profile: 55 ºC for 60 minutes, 70 ºC for 15 minutes, and 4 ºC until samples were removed. Finally, 1 uL of RNase H (Invitrogen, USA) was added for incubation at 37 ºC for 20 minutes to remove RNA. The sample could be stored at -20 ºC before subsequent steps.

The cDNA amplicon generation was performed with Phusion polymerase (NEB, USA). The 50 uL reaction mixture was prepared with the following components: 5x High-Fidelity buffer in 10 uL, 10mM dNTP in 1 uL, 10 uM FR3 degenerate forward

primer (5'-AAGMRGACAYRGCYRTSTATTACTG-3') in 5 uL, 10 uM IgM reverse primer (5'-AACGGGGAATTCTCACAGGAGAC-3') in 5 uL, first-strand cDNA template in 10 uL, 2 units/uL of Phusion polymerase in 1 uL, and DI water to top off final volume to 50 uL. The reaction mixture was then subjected to thermocycling with the following temperature profile: 1 cycle (98 ºC, 30 seconds), 4 cycles (98 ºC, 10 seconds; 55 ºC, 30 seconds; 72 ºC, 30 seconds), 4 cycles (98 ºC, 10 seconds; 50 ºC, 30 seconds; 72 ºC, 30 seconds), 4 cycles (98 ºC, 10 seconds; 45 ºC, 30 seconds; 72 ºC, 30 seconds), 6 cycles (98 ºC, 10 seconds; 50 ºC, 30 seconds; 72 ºC, 30 seconds), 1 cycle (72 ºC, 5 minutes), 4 ºC until samples were removed. The reaction products could then be size selected using either common gel extraction method or Agencourt AMPure XP beads size selection described below. The samples could be stored at -20 ºC until subsequent steps.

Agencourt AMPure XP beads size selection

Agencourt AMPure XP beads (Beckman Coulter, IN, USA) are solid phase reversible immobilized beads that retain various sizes of DNA based on the flocculation condition. By increasing the volume ratio between the AMPure XP beads to the DNA samples, the DNA size retention can be extended to shorter fragments. In general, the following beads to sample ratio can provide the corresponding size retention: 0.4X for >1.2kbps; 0.6X for >900bps; 0.8X for >300bps; 1.2X for >100bps, etc. The experiments conducted in this manuscript used the ratio ranges from 0.5X to 0.85X to cover sizes of 250bps to about 1kbps. The beads should be allowed to reach room temperature before use. The room temperature beads were vortexed with the samples and allowed to incubate at room temperature for 15 minutes and then magnetized for 5 minutes in MagneSphere magnetic stand (Promega, WI, USA). While placing the tube in the magnetic stand, the

supernatant was removed and the remaining beads were washed twice with 80% ethanol. The tube was removed from the stand and allowed to air-dry for about 10 minutes. Then, 33 uL of Tris-EDTA (TE) was added to the beads and vortexed for about 15 seconds. The tube was incubated at room temperature for 5 minutes and then magnetized for another 5 minutes. About 30 uL of the supernatant was transferred to fresh tube and these size-selected samples were ready for subsequent steps or storage at -20 ºC.

Overview of circle sequencing

An overview of the circle sequencing process is summarized in Figure 27. The process mainly consisted of four main modules: Template circularization, Rolling circle amplification, Next generation sequencing and Bioinformatics analysis.

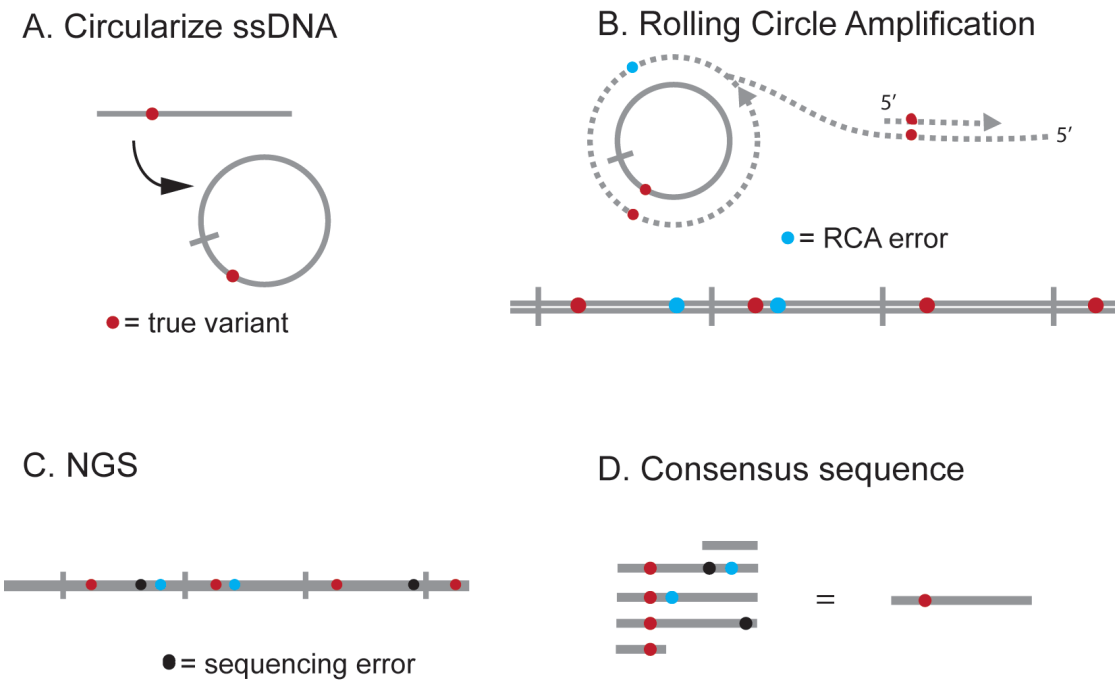

Figure 27: Overview of circle sequencing

Template circularization

The amplicon should be phosphorylated at the 5' end before circularization. Details of phosphorylation can be found in most kits. Briefly, first, the amplicon was made into ssDNA by denaturation at 95 ºC for 15 minutes. Denatured amplicon was snap-freezed in liquid nitrogen for 5 minutes and subsequently allowed to thaw on ice. While thawing, the following CircLigase II (Epicentre, WI, USA) reaction mixture was prepared: 10x CircLigase II reaction buffer in 2 uL, 50mM $MnCl_2$ in 1 uL, 5 pmol ssDNA in <16 uL, 100U CircLigase II in 1uL, where the final reaction volume was at 20 uL. The reaction mixture was incubated at 60 ºC for 1 hour. Removal of non-circularized ssDNA or dsDNA was completed by adding 1 uL of Exonuclease I (NEB, USA) and 0.5 uL of Exonuclease III (NEB, USA) to each reaction and incubated at 37 ºC for 1 hour. PCR cleanup was performed and circularized products were eluted with 20 uL DI water. The products could then be stored at -20 ºC.

Rolling circle amplification

The ability for Bacteriophage Phi29 DNA polymerase (NEB, USA) to strand displace facilitated the generation of tandem repeats from the same circularized template. However, due to its 3'-to-5' exonuclease activity, exonuclease-resistant primer containing phosphorothioate bonds are necessary to maintain the efficiency. First, the primers described in the cDNA generation step with phosphorothioate bonds were allowed to anneal to the circularized template in a 20 uL mixture. The mixture contained the followings: 2X Annealing buffer in 10 uL [10 mM Tris pH 8, 50 mM NaCl, 1mM EDTA], 10uM of both forward and reverse exo-resistant primers in 1 uL, circularized

template (1-100 ng) in <9 uL, and finally, DI water diluting the mixture to a 20 uL total volume. The mixture was allowed to incubate at 95 ºC for 5 minutes and cooled to 4 ºC. Then, the following components were added to the 20 uL mixture: 5 uL of 10x Phi29 DNA polymerase reaction buffer, 1 uL of 100x BSA, 1 uL of 10mM dNTP, 0.2 Units of Inorganic pyrophosphate (NEB, USA), 10 Units of Uracil-DNA glycosylase (NEB, USA), 16 Units of Formamidopyrimidine-DNA Glycosylase (NEB, USA), 2 uL of Phi29 polymerase (NEB, USA), and finally, DI water added to a final reaction volume of 50 uL. The reaction mixture was allowed to incubate at 30 ºC for 3 hours and then heat-inactivated at 65 ºC for 10 minutes. Finally, the rolling circle products were subjected to ethanol precipitation.

Ethanol precipitation

The rolling circle products were mixed with a 1/10 volume of 3 M pH 5.2 sodium acetate. After thoroughly mixing, 3X volume of 100% ethanol was added and the mixture was placed at -80 ºC freezer overnight. The mixture was then spun at max speed for 30 minutes at 4 ºC to pellet the DNA. The supernatant was discarded and 500 uL of fresh 70% ethanol was added and vortexed for about 5 seconds. The resuspended mixture was then spun at max speed for 15 minutes at 4 ºC. The supernatant was carefully removed and the DNA pellet was allowed to air-dry for about 10 minutes. Then, the DNA pellet was resuspended with 130 uL of Tris-EDTA (TE). The resuspended DNA products were then ready either for storage at -20 ºC or subsequent steps.

Shearing of the rolling circle products

Covaris S2 sonicator (Covaris, MA, USA) was used to shear the rolling circle to a range of predefined sizes. The sonicator preparation procedure can be referred on the manufacturer's website and the settings used in the manuscript were as follows. For fragment size of about 1kbps, duty cycle at 2%, intensity at 4, cycle per burst at 200, time at 22 seconds, and mode is Freq Sweeping. For a desired fragment size of about 800bps, duty cycle at 5%, intensity at 3, cycle per burst at 200, time at 25 seconds, and mode is Freq Sweeping. After sonication, the sheared products could be purified with PCR cleanup kit and eluted in 25 uL of DI water. This purified product would be ready for sequencing preparation.

TruSeq sequencing sample preparation

This step was conducted by the Genomic Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin. Briefly, the procedure entails end repair of sheared products, addition of adenine to repaired products, ligation of the TruSeq adaptors, 10 cycles of amplification to amplify successfully ligated products, AMPure XP beads size selection. The end products of the procedure would then be ready for Illumina MiSeq sequencing.

Synthetic IgM control construction

Plasmid containing pre-defined heavy chain IgM sequence was constructed using the Gibson assembly cloning method as described in the manufacturer's manual (NEB, USA). The backbone of the plasmid was derived from the pMAZ360 vector [225] and the IgM sequence construct was generated with synthetic DNA sequences (IDT gBlock,

USA). The plasmid contains an origin of replication from pBR322 as well as an ampicillin selection marker. The synthetic heavy chain IgM is under a T7 promoter allowing generation of the synthetic IgM mRNA. The variable region and the CH1 of the sequence are shown in Figure 28 and the full sequence is shown in the Appendix.

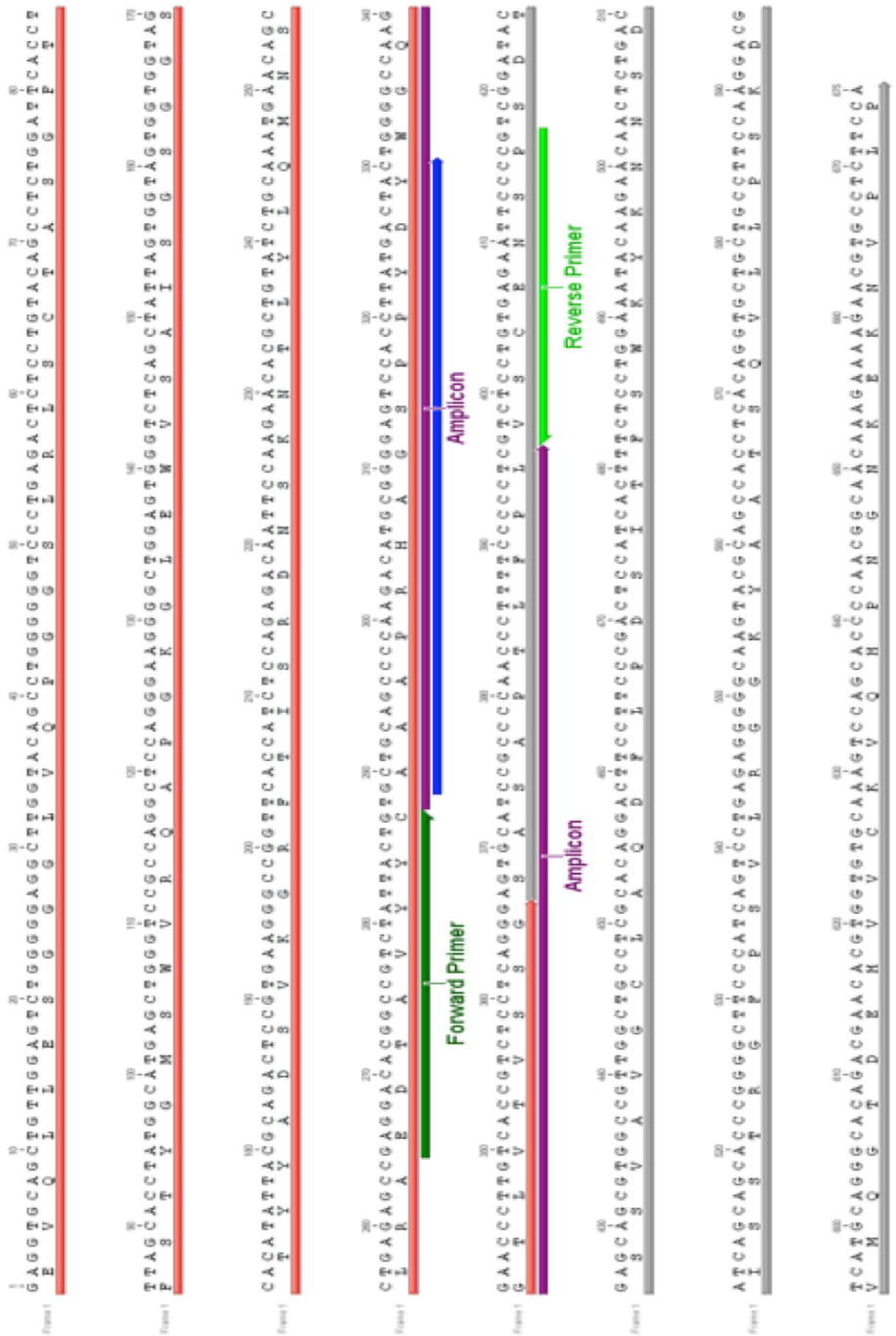

Figure 28: Synthetic IgM construct sequence from variable region to CH1

Note: The orange bar highlights the variable region; the green bars highlight the primers used for amplicon generation; the purple bar highlights the amplicon region; the blue bar highlights the CDRH3 region

Synthetic concatemer IgM control construction

In addition to the synthetic IgM construct, six constructs containing different repeating numbers of the amplicon region (purple-highlight region in Figure 28) were constructed using overlap PCR extension and Gibson assembly cloning method. These additional constructs included amplicon region containing between 2-repeats (two-mer) to 7-repeats (seven-mer). The purpose of these specific constructs was to produce tightly length-restricted sequencing templates for identifying an optimal shear length that would be the least affected by PCR-mediated recombination event. PCR-mediated recombination as explained in Lou et al.'s study [224]. It is largely believed to be due to incompletion of PCR extension in each cycle of the cluster generation step for the Illumina sequencing platform. Due to the highly repeating nature of the sequencing template, the incomplete product from the previous cycle could complement with another template producing a much shorter product. After cluster generation in Illumina procedure, a cluster would contain a mix of completely extended products as well as shorter products. Thus, after the first repeat read-through of the cluster, some members of the cluster would signal the next correct repeat while some members of the cluster would signal the adaptor sequence. This mixed signal could occur in any subsequent repeats stochastically resulting to reduction of read base quality score for the duration of the length of the sequencing adaptor.

Transcription of the synthetic IgM construct

AmpliScribe T7 High Yield Transcription Kit (Epicentre, WI, USA) was used to generate the RNA from the synthetic IgM construct plasmid. Briefly, the following

reaction mixture was prepared at room temperature: about 200 ng of the IgM plasmid, 10x AmpliScribe T7 reaction buffer in 2 uL, 100mM ATP, CTP, GTP, UTP in 1.5 uL each, 100mM DTT in 2uL, RiboGuard RNase Inhibitor in 0.5 uL, AmpliScribe T7 Enzyme Solution in 2 uL, and top off with RNase-free water to 20 uL total volume. The reaction mixture was incubated at 37 ºC for 2 hours and 2 uL of DNaseI was added to the mixture for another incubation at 37 ºC for 15 minutes. The reaction product was then purified following the total RNA purification process described above. This IgM RNA product was used as the control for assessing the errors generated during the sequencing process.

Bioinformatics analysis: oriented reads processing

Tandem repeats from the circle sequencing are processed *in silico* to assemble the "oriented" read containing the consensus sequence. The bioinformatics pipeline described in the Lou et al.'s study [224] was used to process the sequencing data in this manuscript initially to generate oriented reads without filtering of the PCR-mediated recombination contaminated regions. Briefly, the pipeline first removed the read pairs that mapped to the PhiX genome. Then, the read pairs that contained premature adapter sequences were removed. After these filtering steps, an optimal periodicity was determined by scanning the sequence for the fraction of base-match in a range of fixed distances. Since tandem repeats should be theoretically a 1:1 copy of each other, the fixed distance that yielded the best fraction of base-match would be the optimal periodicity. Once the periodicity was determined for each pair-end reads, the optimal offset between the reads was determined similarly by finding the best-fixed distance that could yield the best base-match fraction. Any pair-end reads without strong periodicity was discarded. With the

periodicity and offset, alignment of the repeats could yield consensus read assembled from independent sequencing events. Thus, the quality score was extended beyond the limit of phred-score Q40. The consensus reads that were adjusted to reflect the primer positions were referred as the oriented reads.

Bioinformatics analysis: seed-based processing

As cautioned by Lou et al. [224] that PCR-mediated recombination has significant impact on the quality of the repeated region of the reads; hence, in order to rescue usable repeats from regions "polluted" with PCR-mediated recombination events, a seed-based processing method was utilized to filter problematic region while reconstructing consensus reads only based on high quality repeating regions. The algorithm for seed-based processing is similar to the Inchworm algorithm in the Trinity *de novo* transcriptome assembly method [226]. A schematic diagram summarizing the process using an example of k-mer with size 4 is shown in Figure 29.

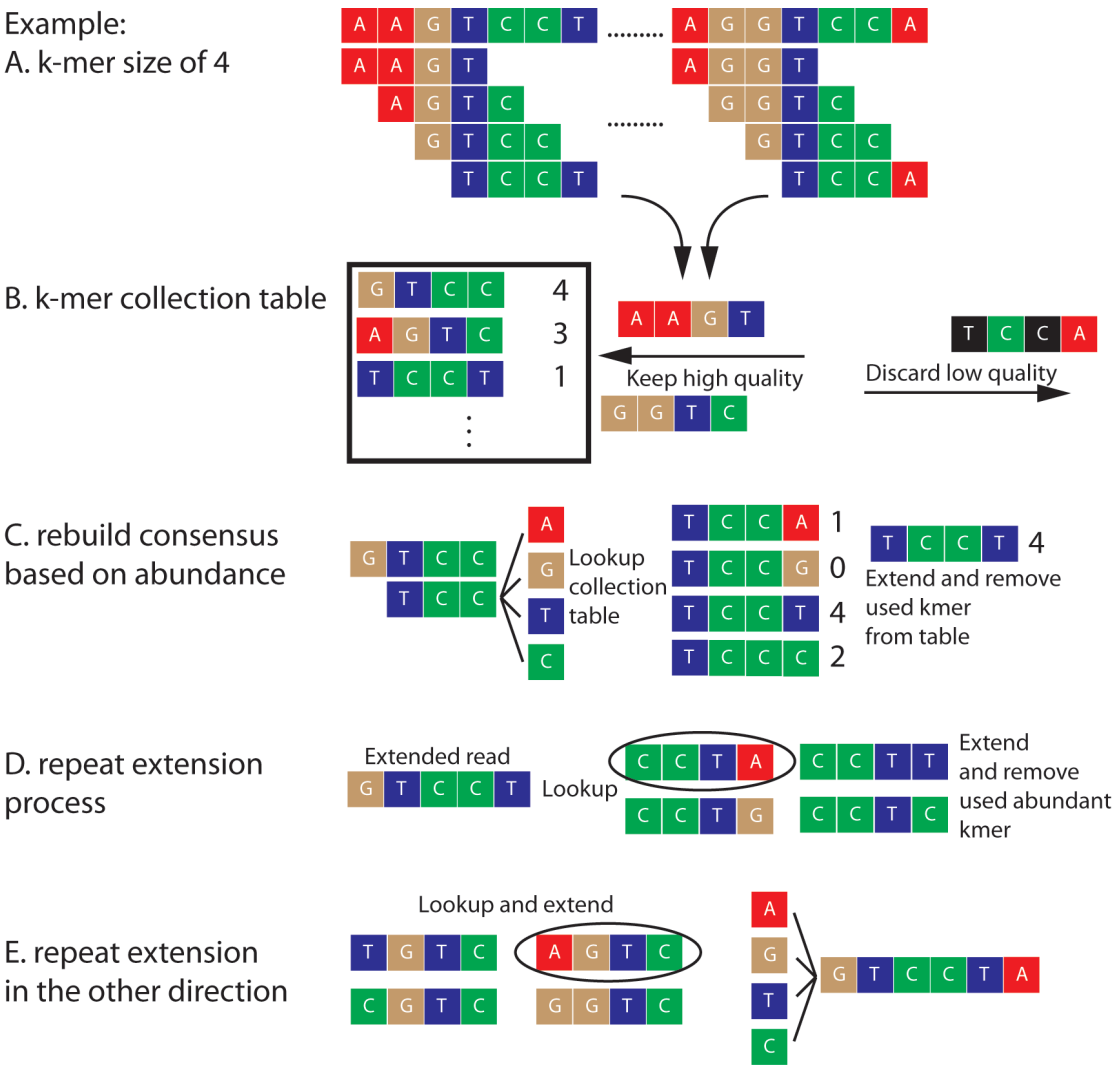

Figure 29: Overview of seed-based processing with an example of size 4 k-mer

Briefly, both the R1 and the reverse complemented R2 reads were broken into a collection of k-mers where k = 9 bases. The k-mers were identified by scanning the window a base apart with width of 9 bases across the whole R1 or reverse complemented R2 reads and the k-mers were quality checked to ensure 80% of the bases contained Q30 or above. Any k-mers that could not satisfy the quality check would not be included in

the collection. When the same satisfactory k-mer was identified, their quality scores per base were expanded by first converting scores back to probability and then the probabilities were multiplied before converting back to the quality score. After ranking the k-mers by abundance, the most abundant k-mer was used as the seed for rebuilding the consensus sequence. Sequence extension to the right began with using the right most k-1 bases from the seed or the extended read. To form k-mer for the extension, the k-1 bases were extended with four possibilities of A, T, G, or, C. By identifying from the collection table the most abundant k-mer among these four possibilities, the read was extended with the base that made the k-mer the most abundant from the collection table. The used k-mer was removed from the collection table preventing it from later consideration. In case of tie abundance, branching off events occurred and the tied counts were added in the next extension until a global maximum abundance was resolved leading to the respective branch kept as the extension result. When the sequence could no longer be extended to the right, the same exercise was repeated to the left until no k-mer could be identified to the left. The resulting extended sequence would have by now removed PCR-mediated recombination events as well as reconstructed the consensus.

Bioinformatics analysis: error rate measurement

The error rate was measured after aligning the determined consensus sequence to the known sequence of the synthetic IgM control construct using Biopython global alignment default settings. Since the sequence of the input was known *a priori*, error rate could then be measured by aligning the sequencing results to the known input sequence. The error rate was calculated by counting the total number of mismatch bases from the control over the total number of bases read.

Bioinformatics analysis: quality score threshold determination

Receiver operating characteristics (ROC) analysis was used to determine the optimal expanded quality score that balances sensitivity and specificity for reliable base detection. Data generated from the synthetic IgM control experiment was used to perform this analysis. Briefly, a list of truly matching bases with their quality scores and a list of truly mismatching bases with their quality scores were collected from each sample (10 for conventional and 10 for circle sequencing-filtered). Quality score threshold ranging from 1 to 93 was used to specify reliable base. Contingency table for each quality score could be constructed to calculate the True Positive Rate (number of truly matching base specified as reliable base over all truly matching bases) and the False Positive Rate (number of truly mismatching base specified as reliable base over all truly mismatching bases). These pairs of FPR and TPR were plotted as scattered plot. The quality score that maintained the FPR below 10% was used to analyze the human naïve B cells samples.

Bioinformatics analysis: CDRH3 identification

Identification of the CDRH3 was done by extracting sequences in between the following motifs:

TATTACTG[ACTG] (…) TGGGG[ATCG][AC][AG][ACTG]GG[ACTG] where bases in the square brackets represent degenerate bases for that position in the motif and the parenthesis represent the CDRH3 sequences to be extracted. Upon extraction of the CDRH3 sequences, its minimum quality score was checked to meet the quality score threshold identified by the ROC analysis (Q30 or Q70). Sequences with at least one base not meeting the threshold was regarded as unconfident reads and discarded.

RESULTS

Error rates for conventional sequencing and circle sequencing

The synthetically constructed IgM control template was used in 20 samples. 10 of the 20 samples were prepared using conventional sequencing (no tandem repeats generated) while the second half were prepared using the circle sequencing method described above. All samples were prepared and submitted for NGS MiSeq 2x250 in 5 independent sequencing runs. A total of about 9 million sequencing reads were generated across all runs. The erroneous bases introduced by each method were measured by aligning the consensus reads to the known IgM control sequence. For each sample, an error rate was calculated by finding the fraction of total number of erroneous bases over total number of bases being analyzed. Unlike the conventional sequencing samples, the circle sequencing samples required bioinformatics processing to reconstruct the consensus reads as described in the method section. Two different bioinformatics processing methods were used on the circle sequencing samples where the first method followed the Lou et al.'s scheme [224] without filtering out problematic region caused by PCR recombination (denoted as CircleSeq samples); the second method followed the seed-based processing where the problematic region is filtered (denoted as CircleSeq-filtered samples). The error rates for each sequencing methods and each bioinformatics processing methods are shown in Figure 30.

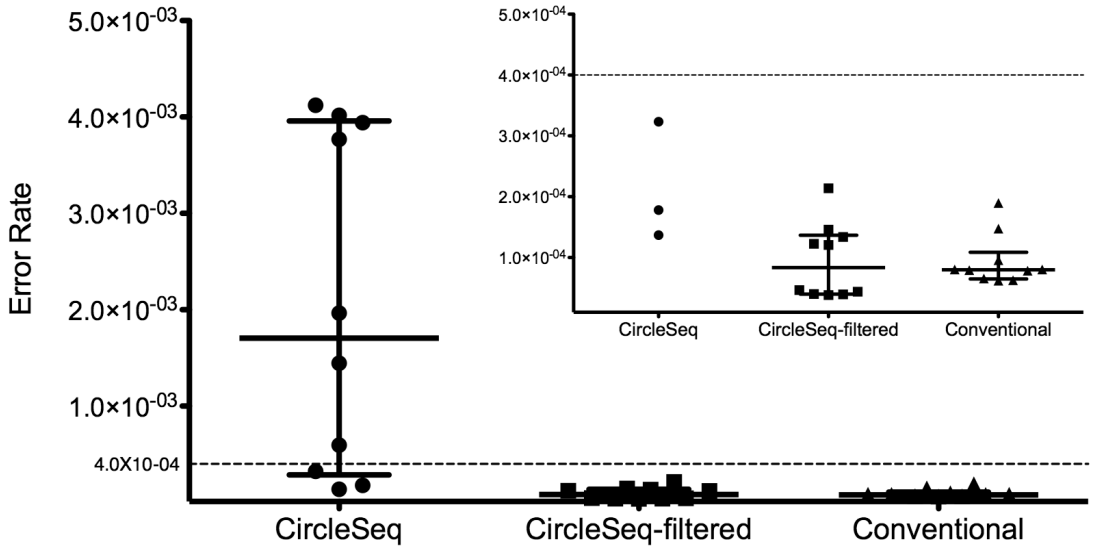

Figure 30: Error rate for the different sequencing and bioinformatics processing methods

Note: CircleSeq is the circle sequencing samples without PCR-mediated recombination filtering. CircleSeq-filtered is the circle sequencing samples with PCR-mediated recombination filtering. Conventional is the conventional sequencing samples. The figure inset has an appropriately scaled y-axis to better show the lower error rate samples.

As seen in Figure 30, without proper filtering of problematic region, the circle sequencing method reported a median error rate of about $1.7x10^{-3}$ error/base. More importantly, non-filtered reads resulted in a wide variation between sequencing runs; the later sequencing runs of the circle sequencing samples reported lower variation and error rates. On the other hand, the conventional sequencing method and the circle sequencing method with filtering both obtained a median error rate of about $8x10^{-5}$ error/base. Specifically, the conventional sequencing was a more robust process in that the samples were consistently obtaining an error rate of $8x10^{-5}$ error/base while the circle sequencing

with filtering samples contained a wider spread. This again could be due to the batch-dependent variation observed from the circle sequencing samples. Additional characterization of the types of mutation revealed that transversion was the predominant cause of mismatches while there appeared to be mutational hotspots throughout the sequence as seen in Figure 31.

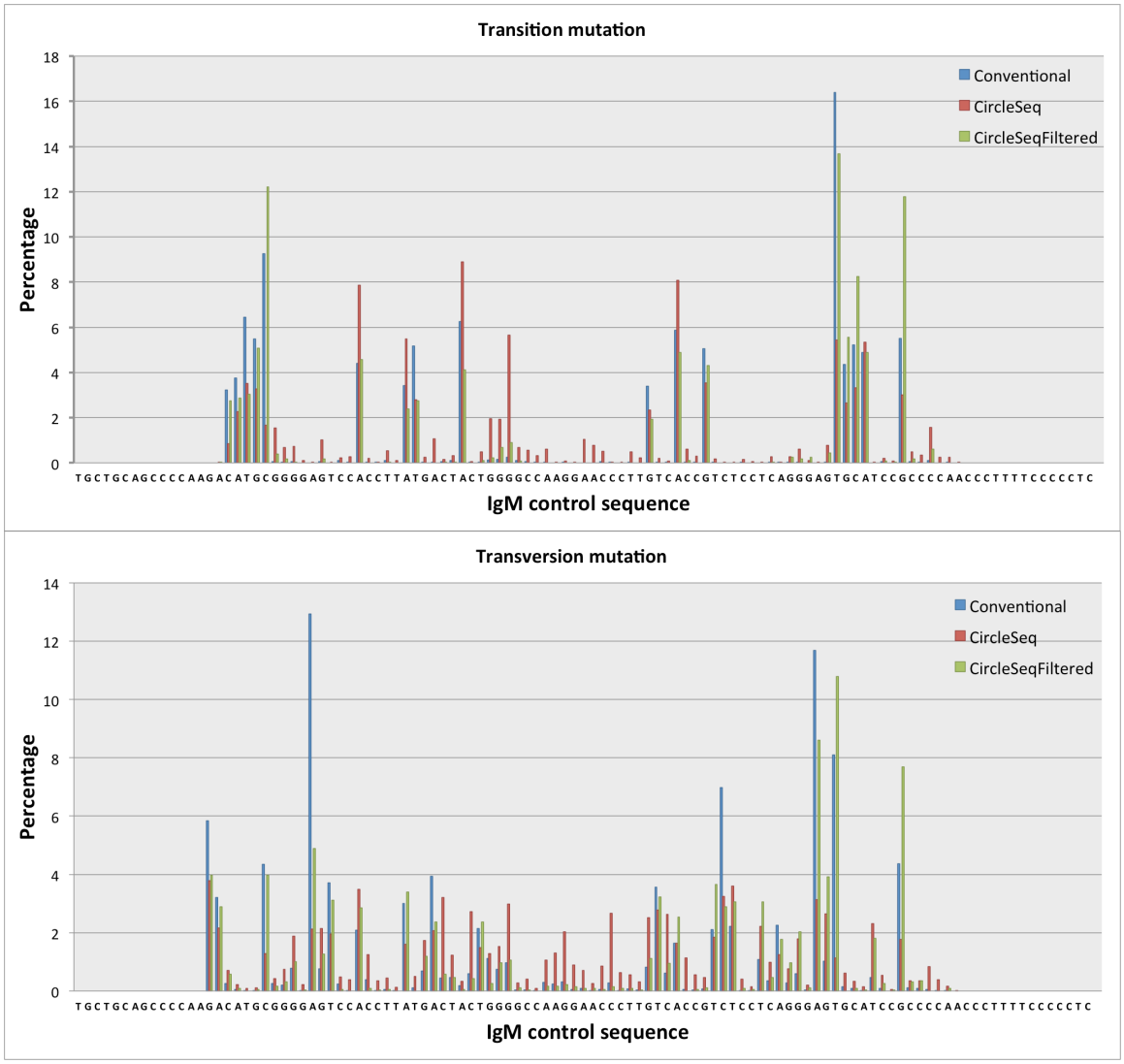

Figure 31: Distribution of transition and transversion mutation

In a separate experiment, we used only 2 (as opposed to 18) cycles of PCR amplification of the synthetic IgM control template. This minimally amplified sequence was prepared and submitted for sequencing. The error rate was also found to be around $8x10^{-5}$ error/base indicating that the lower bound error rate of the process was most likely reached. This assumption is further supported by the fact that cluster generation requires 35 cycles of amplification with Bst polymerase (error rate $1.5x10^{-5}$ error/base) [215] and the reverse transcriptase error rate is at about $4.5x10^{-5}$ error/base [227] during the first strand synthesis process.

The severity of the PCR-mediated recombination can be visualized using a complexity plot similar to the one described in the Lou et al.'s study. A complexity plot generated with a synthetically constructed concatamer two-mer sample is shown in Figure 32 to illustrate the problematic region issue that pertains to tandem repeats samples in MiSeq sequencing.

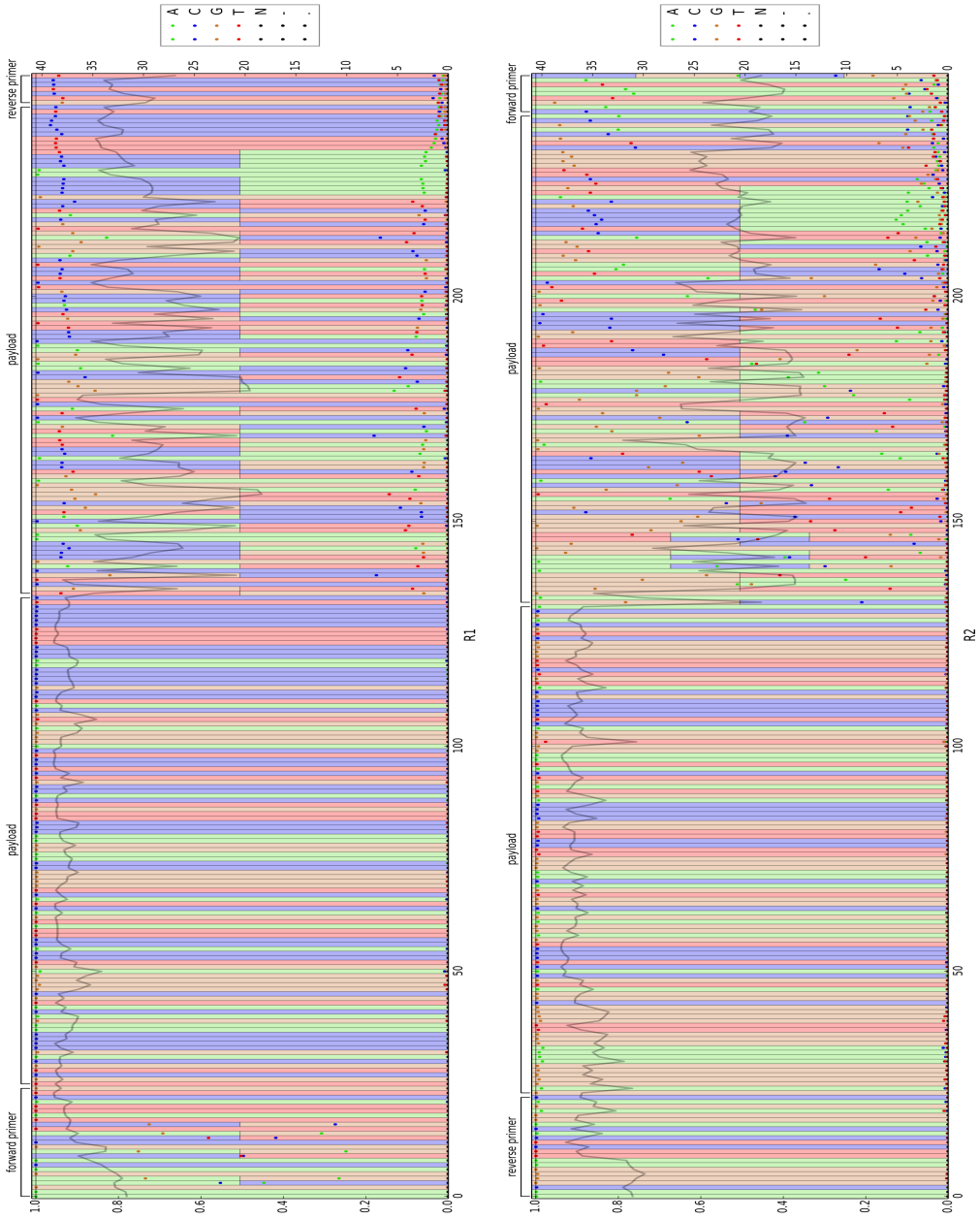

Figure 32: Complexity plot for the R1 and R2 reads of one circle sequencing sample

Note: the left panel is R1 read and the right panel is R2 read; payload is the region of interest as well as the repeating unit; the left y-axis is fraction; the right y-axis is quality score; for each position, the dots represent the fraction of the bases reported from the sequencing run, the shaded background represent the supposed base for that position; the split shading represent degeneracy or sequencing adaptor region

The complexity plot is reporting the relationship between quality score of a sequence to the different distribution of bases reported for each position of the 250 bps from each of R1 and R2 reads in a run. The sequenced samples were the two-mer samples containing exactly one forward degenerate primer region, two payload regions (IgM CDRH3-FR4 contorl sequence), and one reverse primer region. The samples were prepared using the conventional method of synthetically designed two-mer tandem repeats of the IgM control sequence (Gibson assembly cloning described above). Hence, this sample should have the least amount of extra confounding preparation steps. The plot indicates that the read quality remained high until the first repeat was encountered. Moreover, during the first repeat, a large portion of the bases reported matched the control sequence. However, the error rate significantly increased once the second tandem repeat was encoutered. With respect to this repeat, the quality score dropped to as low as 15 (equivalent to 0.032 error/base). It was apparent that within the tandem repeat, there was a strong emergence of sequencing adaptor sequences denoted in the splitted shading background within the tandem repeat section in Figure 32. This drop in quality and emergence of adaptor sequences was due to PCR-mediated recombination cautioned by Lou et al. Incomplete extension during cluster generation and homologous repeats facilitated shorter reads within a cluster. Hence, in extreme severtity, a cluster might be hijacked with reads containing less repeats while in moderate case, the read quality would drop drastically during the repeats. Therefore, it is necessary to filter the circle sequencing reads to prune unconfident sections of the reads caused by PCR-medaited recombination as evidence in Figure 30 confirmed the case.

Effects of PCR-mediated recombination on different template lengths

As mentioned above, circle sequencing is prone to PCR-mediated recombination leading to a degradation of high quality reads. In order to find an optimal template length, containing tandem repeats, that is the least affected by PCR-mediated recombination we created synthetic control sequences. Each control contained different numbers of tandem repeat of the known IgM CDR3-FR4 control sequence (payload). These concatemers were prepared using the conventional sequencing method. The sequencing reads were then processed using Lou et al.'s bioinformatics processing scheme to recover the consensus reads. These consensus reads were then aligned to the known IgM CDR3-FR4 control sequence to determine the error rate as an indirect measure of PCR-mediated recombination effects.

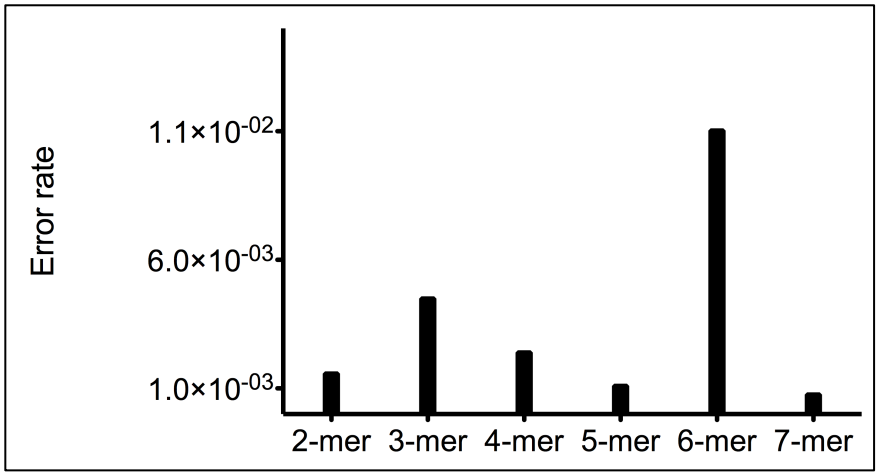

Figure 33: Error rate for the different concatemers

Theorectically, the longer template containing more repeats (i.e. 7-mer) should be more prone to PCR-mediated recombination as there are more "entry points" for homologous binding. The longer length should also favor incomplete cluster generation.

Despite these speculations, as seen in Figure 33 except for the outlier 6-mer sample, the longer concatemers (5-mer and 7-mer) have similar error rate around $1x10^{-3}$ error/base to that of 2-mer. Based on this evidence alone, it is still unclear whether there is an optimal length that is the least affected by PCR-mediated recombination. However, at the very least, it is obvious that PCR-mediated recombination will affect any tandem repeat containing template as the error rate for these samples were about an order of magnitude higher than those without tandem repeats (i.e. conventional sequencing sample in Figure 30).

Determination of quality score threshold for improved sensitivity to true variant

Although circle sequencing cannot outperform conventional sequencing in terms of the overall error rate, it does maintain its advantage where bases that are confirmed by redundancy have an expanded quality score. This expanded quality score allows for a greater confidence in the detection of a true variant at a specific position. The sensitivity of a method in detecting true variant is best visualized using a receiver operating characteristic (ROC) curve. This curve is originally used to illustrate the performance of a binary screening system where it plots the sensitivity against the non-specificity as the screening threshold changes. In this case, the match and mismatch information together with its respective quality score from all the synthetic IgM control sequencing experiments were collected accumulatively to determine the sensitivity and the non-specificity. The sensitivity (or True Positive Rate) is the fraction of matching bases above the quality score threshold from all the matching bases. The non-specificity (or False Positive Rate) is the fraction of mismatching bases above the quality score threshold from all the mismatching bases. These pairs of values are plotted over the range of the

expanded quality score (i.e. Q1 to Q93). The ROC plot for sequencing samples derived from the conventional sequencing and the filtered circle sequencing is shown in Figure 34.

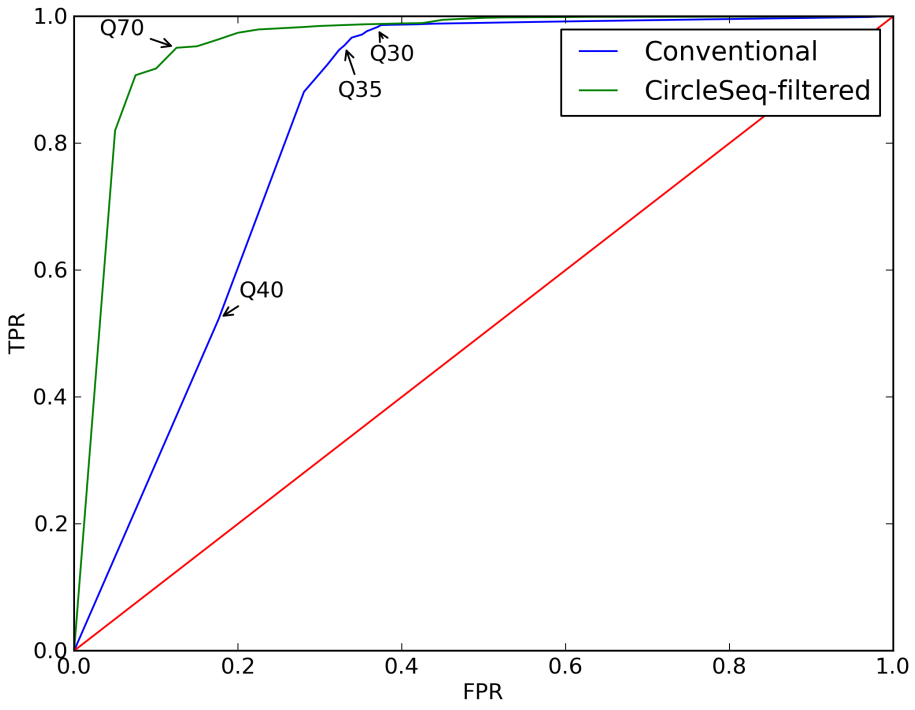

Figure 34: Receiver operating characteristic (ROC) curve for the conventional sequencing and the circle sequencing PCR-mediated recombination filtered methods

Note: TPR stands for True Positive Rate and FPR stands for False Positive Rate; Several quality score threshold points were annotated to illustrate quality scores for normal practice and the optimal quality score for the filtered circle sequencing process

As important reference on a ROC curve, the diagonal line and the upper-left hand corner represent two extreme cases. The diagonal line reflects that the varying threshold

parameter has random sensitivity/specificity perfomance whereas the corner reflects perfect sensitivity/specificity performance. As seen in Figure 34, the filtered circle sequencing method outperformed the conventional sequencing method in its improved sensitvity/specificity performance. More specifically, the highest possible quality score (Q40) in the conventional sequencing gave about 0.52 TPR and 0.18 FPR. This means that using >Q40 threshold, one can expect to confidently retain 52% sequencing-error-free bases while bearing 18% risk of mistrusting a base as error-free. On the other hand, with the expanded quality score in the filtered circle sequencing method, Q70 threshold gave about 0.95 TPR and 0.13 FPR which is optimally balanced to yield the highest possible sensitivity while managing the non-specificity. After obtaining this important threshold parameter, detection of true variant becomes much more trustworthy, especially when incorporated into a clustering algorithm.

Measurement of human naïve B cells with the different sequencing methods

In the absence of a diverse pre-defined repertoire standard, it is quite difficult to quantify the true variant detection performance from each sequencing method. However, as an approximation, measurement of a pre-defined number of human naïve B cells could still provide indicative benchmark to judge relative performance. It is empirically estimated that the naïve B cell diversity is about 1-10 million clones [69]–[71]. Therefore, sampling a small fraction of this vast diversity, say 200,000, should give rise to a maximum clonal diversity of 200,000. Using this hypothesis, any acceptable sequencing method should reports clonal diversity within reasonable range of 200,000.

To test this hypothesis, 200,000 naïve B cells from the same individual were collected, counted, and extracted for the total RNA. After 18 cycles of cDNA amplification, half of the amplified library, the conventional sequencing sample, was submitted for MiSeq 2x250. The remaining half was prepared with the circle sequencing method. Generated tandem repeat products were sheared at length of about 1 kbps prior to MiSeq 2x250 sequencing. After applying the seed-based processing step on the circle sequencing reads, the recovered sequences were about 4 million reads. The conventional sequencing yielded about 7.4 million reads. In order to normalize the number of reads between the samples, 4 million reads were randomly drawn from the conventional sequencing set for the analysis. As mentioned previously, applying a quality score threshold to prune unconfident bases can be an effective way to remove artifactual variants. A Q70 threshold was established to be optimal for the filtered circle sequencing sample. The same threshold is not possible for the conventional sequencing sample; in addition, using a Q40 would filter out most sequences from a conventional sequencing sample. Since Q30 is frequently used as a common quality cutoff, it will be used as threshold here to represent a normal practice.

The unique numbers of CDRH3s generated by both of the sequencing method were higher than the maximum of 200,000 clones. This increased diversity suggests that both methods contained artifactual variants. From the conventional sequencing sample (Q30), there were 1,226,653 unique CDRH3s. On the other hand, the filtered circle sequencing sample (Q70) was analyzed to report 254,287 unique CDRH3s. The inflated diversity seemed more severe in the sample prepared with conventional sequencing method even with the commonly used quality score cutoff. Applying the circle sequencing method and the Q70 cutoff seemed to have mitigated the issue better than the

conventional sequencing method. Table 11 presents the top 30th abundant CDRH3s from each sequencing method sample.

Conventional	Frequency	CircleSeq-filtered	Frequency
ARDETSVGPITGTTPSYYFDY	0.00043	ARAYSTGPQESYFAY	0.000743
ARAYSTGPQESYFAY	0.000325	AKGYDSSGYYFDY	0.000254
ARAWGYDSRESFYSY	0.000183	ARAWGYDSRESFYSY	0.000212
ARGLLSAASLNWFDP	0.000144	ARGLLSAASLNWFDP	0.000198
AREDGAVAGPYTGGMDV	0.000118	ARDETSVGPITGTTPSYYFDY	0.000129
ARGLYCGGDCYPGSLGPDYYYDMDV	0.000108	ARGGRLITMLRGVRNYFDY	0.000122
ASEGGDYYGAGRMLGGDY	0.000103	VKDDTPLLYYYGSGSYYSY	0.000118
ARGRVGRLRVAVAGTGQGPRYFDY	0.000103	AKDGPDYGGNPFDY	0.000115
ARGAYYYDTSGYSGAFDI	0.000098	ARDLGYYDSSGYNL	0.000104
ARGTLEGLKKASWRRLWSHGFDP	0.000097	ARDAHSSGLDAFDI	0.000101
ARGRSEYCSGGSCYSGRKNYYFDY	0.000096	AKEGIVLIVYATSFDL	0.000097
ARESHDTGWFDY	0.000082	AKVNIVLIIYASGFDY	0.000094
ARGAPSSITMVRGVYYYLDY	0.00008	ARNYDTSAYYYYF	0.000094
ARGAYCGGDCYPGLIPHWNFDL	0.000078	ARGPDSSNFYYFY	0.00009
ASAGYYYDSSGYYWIAAFDY	0.000077	ARVAYIYYDFWSGYYYFDY	0.000087
ARVRDGSGSYSWYGMDV	0.000076	ARSGYDSSGYGRADVN	0.000087
AGGIAAAGTSLNWFDP	0.000075	ARVGPVLGSYRYIDY	0.000083
ARVFGSGFPPPSDAFDI	0.000072	ARDYYDSNA	0.00008
ARDATIIYSGSYLANYYYGMDV	0.000071	TSPHYDSSDINDY	0.00008
TTDGPGLRFLEWLSYYYYYYGMDV	0.000069	AKDALVDKYSGSYSDY	0.00008
ARGSGDQILTAFYYYY	0.000064	ARATGILTGYYD	0.00008
AKDGPDYGGNPFDY	0.000064	ASWYYYDSSGYYLFY	0.00008
ARIKFRGLIGTTKYYYYGMDV	0.000062	ARDFDY	0.00008
ARDDIAADGEWFDP	0.00006	ARSTSSWEDWLDS	0.000076
ARGLIAAAGISYYYYAMDV	0.000057	VREFVRGVIISYFDS	0.000076
ARGLVAAAGISYYYYGMDV	0.000057	ARDSTTFGAFDI	0.000073
ARDPGSGYYYVSRWRAFDI	0.000057	ARDWHDISGYFEY	0.000073
AKEGIVLIVYATSFDL	0.000056	ARHSSTNFFDY	0.000073
ARGPDSSNFYYFY	0.000056	ARATVETNWLDP	0.000073
ARGGRLITMLRGVRNYFDY	0.000056	ARIKFRGLIGTTKYYYGMDV	0.000069

Table 11: Top 30th abundant CDRH3s amino acids sequences from the conventional sequencing and the filtered circle sequencing method

Note: the shaded CDRH3s are the ones shared in both samples

As observed in Table 11, not all highly abundant CDRH3s were shared between the samples. Only 8 out of 30 CDRH3s were shared indicating the additional steps involved in the circle sequencing method might have imposed bias affecting the abundance. This was expected by reasons that circularization efficiency was reported to be length dependent [228], [229]. This could explain why there was a descrepancy in the members of the top 30th abundant CDRH3s between the two samples. This trend persists for the total length distribution as seen in Figure 35. The CDRH3s in the filtered circle sequencing sample (Q70) were shorter (mean at 14.5 a.a.) as compared to the conventional sequencing sample (Q30) (mean at 17.5 a.a.). It is important to note that the results generated by the filtered circle sequencing process (Q70) followed more closely to the healthy human CDRH3 length distribution with mean near 14 a.a.

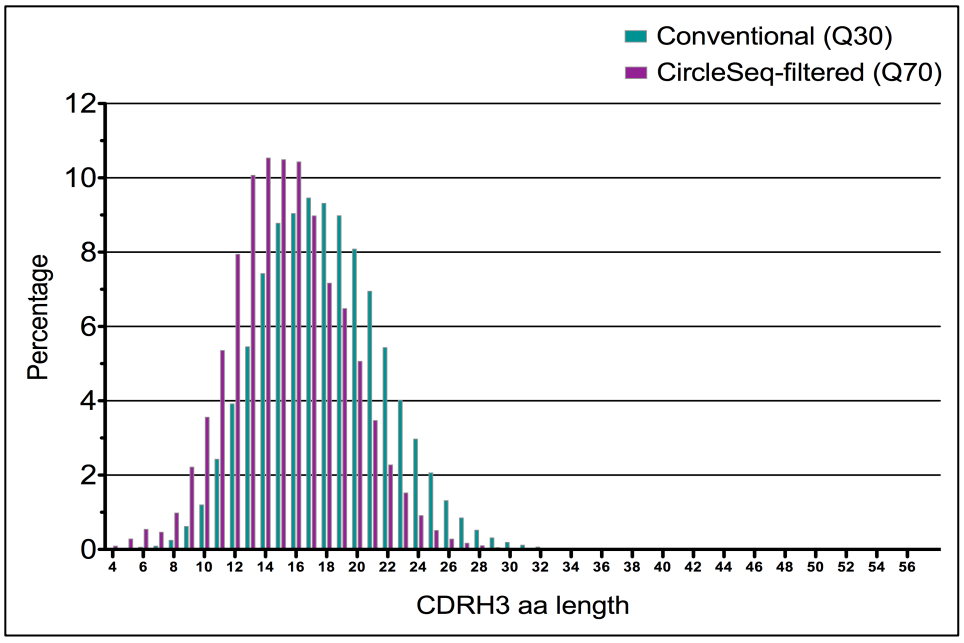

Figure 35: CDRH3 amino acids length distribution between the conventional sequencing and the filtered circle sequencing sample

On the other hand, there appeared to be no difference for the hydrophobicity index [230] across the two methods as shown in Figure 36. The distribution also follows similarly to healthy individual in other reports [18], [210], where the mean is around the neutral (index of zero) and there is a long tail towards the hydrophilic end (negative index). This implied that biases towards base composition might be minimal and not as influencial as the length to the individual method.

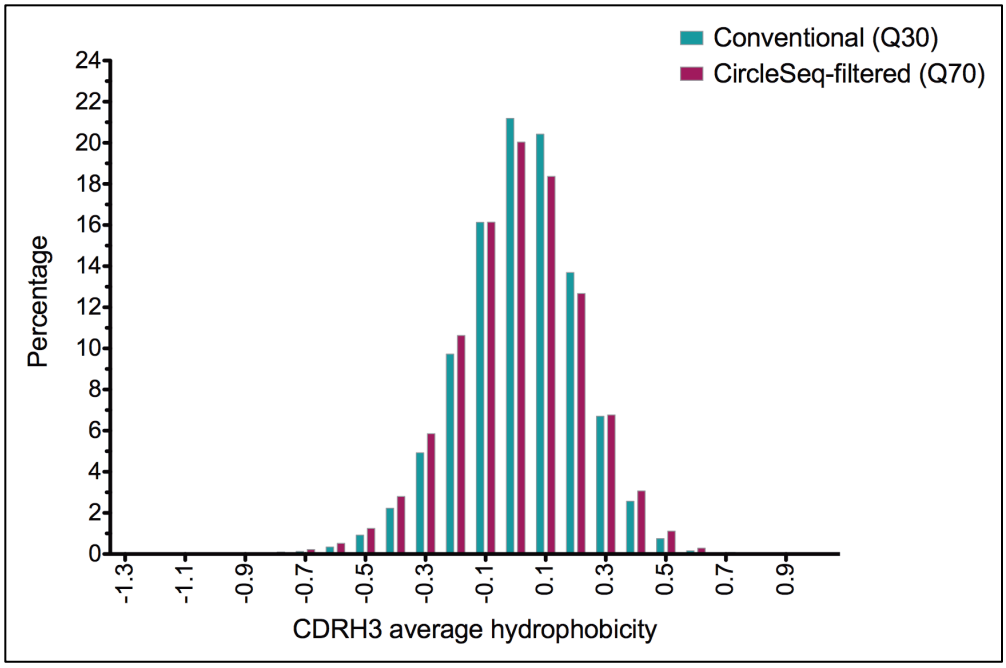

Figure 36: CDRH3 average hydrophobicity between the conventional sequencing and the filtered circle sequencing sample

DISCUSSION

The ability to detect a true sequence variant is entirely dependent on the ability to reduce sequence noise introduced by sequencing error. Thus, there is a great demand for

a silver lining solution that mitigates the overall error rate imparted by next generation sequencing to the antibody repertoires. The circle sequencing method described in Lou et al.'s study [224] provided an alternative approach to barcoding method for improving the sensitivity and detection of true sequence variants. We have adopted and characterized the circle sequencing method to analyzing the diversity of the CDRH3 sequence found within the naïve B cell antibody repertoire. Although the overall error rate for the circle sequencing method cannot outperform the conventional sequencing method, it did perform equivalently for an error rate at about 10^{-5} error/base given proper filtering of problematic sequence regions. In theory, the circle sequencing method should warrant ultra-low error rate but the negative impact by PCR-mediated recombination might have overshadowed the performance. To prevent PCR-mediated recombination issue, it would be necessary to modify the proprietary cluster generation process of Illumina sequencing. Particularly, to guarantee completion of partial sequence extension, we would need to change the polymerase or lengthen the extension time. Also, although not yet optimized for this current project, the initial number of cDNA amplification cycles should be minimized to still allow sufficient circularization. This could minimize early-on mutations from propagating to the circularization step.

Despite the issues affecting the performance of circle sequencing, the unique benefit provided by circle sequencing is the increase in confidence for each base it sequenced. The ability to expand the quality score from Q40 to Q93 is extremely useful for robust screening of true variants from artifactual variants. Providing the expanded quality score, the circle sequencing method greatly improved the sensitivity of true variant detection. As proof of concept example, 200,000 human naïve B cells CDRH3s were sequenced with the two sequencing methods. Combined with the commonly used

quality cutoff (Q30) and the expanded quality score cutoff (Q70), the results were indicating that both methods reported inflated diversity. However, the circle sequencing method is the least affected and reported the number of unique CDRH3s as closely to the theoretical clonal numbers. Hence, the circle sequencing method with the expanded quality threshold was more effective at removing artifactual CDRH3 variants. Moreover, the CDRH3 repertoire reported by the circle sequencing method resembled closely with publicly reported values. Thus, the benefit of expanded quality score as a method for screening true variant outweighs the moderate error rate of 10^{-5} error/base. It would be recommended to adopt the circle sequencing method along-side the conventional method where one can take advantage of the full-length information in the conventional method and the true CDR3 variant confidence in the circle sequencing method.

Chapter 6: Conclusion and future work

NGS has presented us with an unprecedented opportunity to systematically survey an immunolglobulin repertoire that could lead to discoveries not possible before. For antibody discovery, the use of NGS has helped to circumvent the need for labor intensive and time-consuming hybridoma screening process. Putting the NGS antibody discovery into perspective, it effectively cut down the time for the process from about 8 months to around a month. In addition to the improvement in antibody discovery process, the survey of the antibody repertoire using NGS has also revealed several interesting biological traits in human, rabbit, and other organisms. Given the high throughput nature supported by the NGS technology, we could plausibly confirm the commonly believed public Vλ clone (CDRL3) across different individuals. We also reported that about 20% of the Vλ repertoire from our healthy donors and humanized mice samples were public clone. After some characterizations of the public repertoire found in the samples, we observed in the public clones that about 5 or less nucleotide SHM and lower N/P nucleotide addition could have contributed to a limited diversity leading to better likelihood for public clones to be independently generated in different individuals. Similarly, high throughput survey of the antibody repertoire in rabbit revealed other interesting aspects including the discovery of new rabbit germline genes and the observation of about 23% gene conversion frequency in rabbit.

As with any great technology, NGS brings forth great throughput and depth to systematic DNA analysis but it has shortcoming as well. Random sequencing errors can be implanted to the sequencing reads leading to artifactual variants obscuring the true landscape of an antibody repertoire. We have adopted and characterized the circle

sequencing method to seek its utility in sensitivity improvement for the detection of true antibody variants in a repertoire. We have demonstrated that the advantage of the circle sequencing method is the capability of expanded quality score that extends beyond Q40 up to a maximum of Q93. The expanded quality score allows for a higher resolution screening of the quality for sequences prepared with the circle sequencing method. By applying a Q70 threshold screening, the circle sequencing method was shown to effectively reduce the artifactual variants, of which the conventional sequencing method could not achieve due to its limited quality score capping at Q40.

The NGS technology has thus far paved a great foundation for the advent and throughput in antibody repertoire study. In my opinion, there are several future directions that might be worthwhile to pursue. First, as NGS technology becomes more cost effective, it should be used as common practice to apply NGS on directed evolution experiment like the one described in Ravn et al.'s study [231]. By analyzing the repertoire derived from each round of enrichment, one could potentially identify the enrichment pattern and thus be able to understand the relationship between the ligand and the enrichment pattern. A database of these enrichment patterns might help train the next generation of molecular dynamics prediction software to improve rational design for binding interaction. Secondly, the finding of public clones could be further explored for their practical utility. For example, we could test for the ability of the public Vλ clone to pair with unrelated heavy chains that can still retain specificity. If these widely "pair-able" Vλ clones were identified, they might prove to be good "structural stabilizer" that can pair with any heavy chains lacking native pairing information. Finally, having a more precise quantification and correlation of the antibody transcript abundance to clonal cell count can enhance the practical reach of the antibody repertoire data. This would be

particularly useful for diagnosis or cancerous cell tracking for leukemia and other B cell diseases. However, cell quantification relies on the elucidation of the relationship between transcript level and cell numbers. To obtain a solution with the current NGS technology, a standard calibrator that can report a run-to-run amplification difference might be a critical first step. The calibrator should be spiked-in to each sequencing sample such that a reference of transcript-to-cell ratio can be reliably obtained from each sequencing experiment. Two pieces of information is necessary for the construction of a standard calibrator: a. the relationship between cell numbers and the antibody transcripts; b. careful calibrator sequence design to minimize amplification bias. We have started some preliminary works on identifying the amount of IgM transcripts generated by human naïve B cells using quantitative PCR (qPCR). And, the regression line associating the naïve B cell number to qPCR cycle number (Ct) is shown in Figure 37. Together with the regression line associating the IgM transcript amount in ng to qPCR cycle number (Ct) as shown in Figure 38, we were able to derive a preliminary relationship between the cell numbers to the amount of IgM transcript the cells produce. The equation is: log(RNA in ng) = 0.892 x log(Cell numbers) – 8.07. As a result, it is estimated that 100 human naïve B cells can generate about 1 fg ($1x10^{-12}$ g) of IgM transcript. There is still more work needed to better characterize amplification bias for the construction of the calibrator sequence. Once this critical piece of information is ascertained, the absolute quantification of clonal cells could be adjusted and normalized based on the spiked-in standard calibrator. Despite the many directions that can expand the reach of NGS technology, all in all, NGS technology has thus far granted us an unprecedented ability to study an antibody repertoire in never before seen ultra-high resolution.

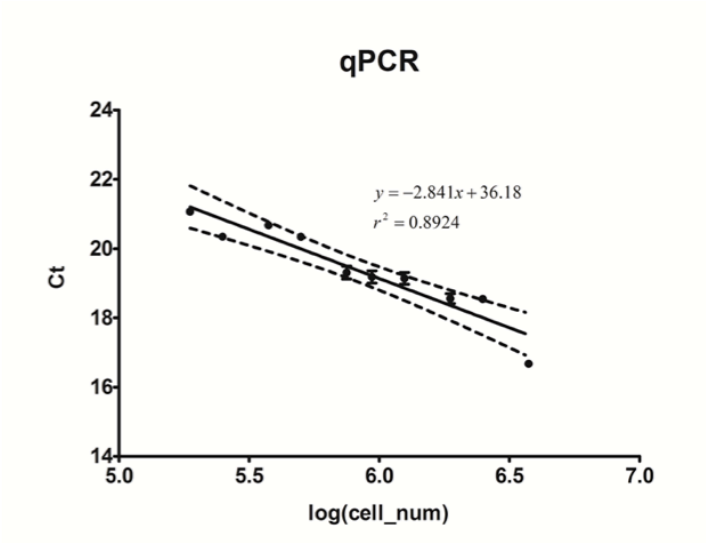

Figure 37: Regression line for Human naïve B cell number to qPCR cycle number (Ct)

Note: the dashed band represents the 95% confident interval for the regression line

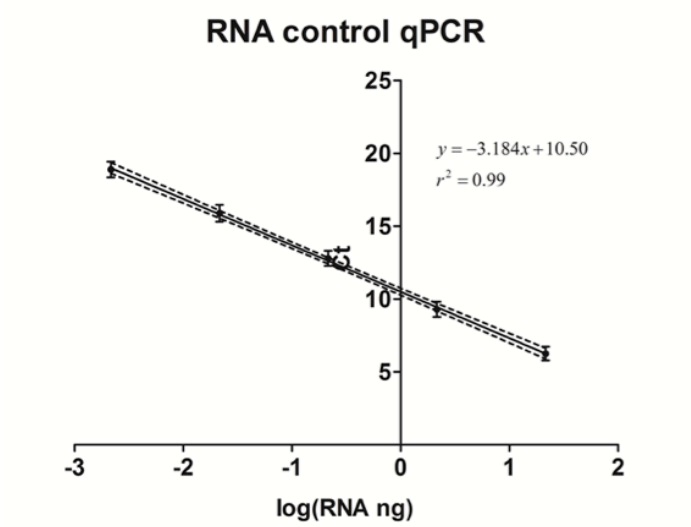

Figure 38: Regression line for IgM transcript in ng to qPCR cycle number (Ct)

Note: the dashed band represents the 95% confident interval for the regression line

Appendices

HUMAN IGM CDNA CONSTRUCT

1 gaggtgcagc tgttggagtc tgggggaggc ttggtacagc ctggggggtc cctgagactc
61 tcctgtacag cctctggatt cacctttagc acctatggca tgagctgggt ccgccaggct
121 ccagggaagg ggctggagtg ggtctcagct attagtggta gtggtggtag cacatattac
181 gcagactccg tgaagggccg gttcaccatc tccagagaca attccaagaa cacgctgtat
241 ctgcaaatga acagcctgag agccgaggac acggccgtct attactgtgc tgcagcccca
301 agacatgcgg ggagtccacc ttatgactac tggggccaag gaacccttgt caccgtctcc
361 tcagggagtg catccgcccc aaccctttc cccctcgtct cctgtgagaa ttccccgtcg
421 gatacgagca gcgtggccgt tggctgcctc gcacaggact tccttcccga ctccatcact
481 ttctcctgga aatacaagaa caactctgac atcagcagca cccggggctt cccatcagtc
541 ctgagagggg gcaagtacgc agccacctca caggtgctgc tgccttccaa ggacgtcatg
601 cagggcacag acgaacacgt ggtgtgcaaa gtccagcacc ccaacggcaa caaagaaaag
661 aacgtgcctc ttccagtgat tgccgagctg cctcccaaag tgagcgtctt cgtcccaccc
721 cgcgacggct tcttcggcaa cccccgcaag tccaagctca tctgccaggc cacgggtttc
781 agtccccggc agattcaggt gtcctggctg cgcgagggga agcaggtggg gtctggcgtc
841 accacggacc aggtgcaggc tgaggccaaa gagtctgggc ccacgaccta caaggtgacc
901 agcacactga ccatcaaaga gagcgactgg ctcagccaga gcatgttcac ctgccgcgtg
961 gatcacaggg gcctgacctt ccagcagaat gcgtcctcca tgtgtgtccc cgatcaagac
1021 acagccatcc gggtcttcgc catcccccca tcctttgcca gcatcttcct caccaagtcc
1081 accaagttga cctgcctggt cacagacctg accacctatg acagcgtgac catctcctgg
1141 acccgccaga atggcgaagc tgtgaaaacc cacaccaaca tctccgagag ccaccccaat
1201 gccactttca gcgccgtggg tgaggccagc atctgcgagg atgactggaa ttccggggag
1261 aggttcacgt gcaccgtgac ccacacagac ctgccctcgc cactgaagca gaccatctcc
1321 cggcccaagg gggtggccct gcacaggccc gatgtctact tgctgccacc agcccgggag
1381 cagctgaacc tgcgggagtc ggccaccatc acgtgcctgg tgacgggctt ctctcccgcg
1441 gacgtcttcg tgcagtggat gcagaggggg cagcccttgt ccccggagaa gtatgtgacc
1501 agcgccccaa tgcctgagcc ccaggcccca ggccggtact tcgcccacag catcctgacc
1561 gtgtccgaag aggaatggaa cacgggggag acctacacct gcgtggtggc ccatgaggcc
1621 ctgcccaaca gggtcaccga gaggaccgtg gacaagtcca ccggtaaacc caccctgtac
1681 aacgtgtccc tggtcatgtc cgacacagct ggcacctgct actgaccctg ctggcctgcc
1741 cacaggctcg gggcggctgg ccgctctgtg tgtgcatgca aactaacccg tgtcaacggg
1801 gtgagatgtt gcatct

CDR3 MOTIF SEARCH (PERL)

```
#! /usr/bin/perl
# This scripts can extract from antibody FASTA file for CDR3 stats, generate top
specified number of CDR3 FASTA for candidate selection

## ask how many top ranked sequences to be reported
print "Enter the number of top ranked FASTA files needed (for consensus): ";
chomp($topnum = <STDIN>);

## Extract FASTA file
$infile=$ARGV[0];
open (InFile, $infile) or die "Error opening $infile !\n";
open (OutFileDist, ">$infile"._Count.txt') or die $!;

@SEQ=();
@label=();
$i = -1;
while ($line=<InFile>) {
 if (($line =~ m/^>/)||($line =~ m/^#/)){
 $newcheck = 1; $i++; $label[$i]=$line;}
 else {
 $line =~ s{[\W\d_]}{}g; chomp($line);
 if ($newcheck == 1) {$SEQ[$i] = $line;}
 else {$SEQ[$i] = $SEQ[$i].$line;}
 $newcheck = 0;
 }
}

$size = @SEQ;
$num50 = 0; $num100 = 0; $num150 = 0; $num200 = 0;
$num250 = 0; $num300 = 0; $num350 = 0; $num400 = 0;
$num401 = 0;
$j = 0;
$lenmax = 0;
$sum = 0;
foreach (@SEQ) {
 $len=length($_);
 if ($len>$lenmax) {$lenmax=$len;}
 $sum=$sum+$len;
 $j++;
 if ($len<=50) {$num50++;}
 elsif ($len<=100) {$num100++;}
```

```
 elsif ($len<=150) {$num150++;}
 elsif ($len<=200) {$num200++;}
 elsif ($len<=250) {$num250++;}
 elsif ($len<=300) {$num300++;}
 elsif ($len<=350) {$num350++;}
 elsif ($len<=400) {$num400++;}
 else {$num401++;}
 }

print OutFileDist "sum = $sum, \t average = ",$sum/$size, "\n";
print OutFileDist "total no. of seq = ",
$num50+$num100+$num150+$num200+$num250+$num300+$num350+$num400+$nu
m401, "\n";
print OutFileDist "Max length = $lenmax \n";
print OutFileDist "\nSeq length\tCount\tPercentage\n";
print OutFileDist "\n 1- 50:\t", $num50, "\t", $num50/$size*100;
print OutFileDist "\n 51-100:\t", $num100, "\t", $num100/$size*100;
print OutFileDist "\n100-150:\t", $num150, "\t", $num150/$size*100;
print OutFileDist "\n151-200:\t", $num200, "\t", $num200/$size*100;
print OutFileDist "\n201-250:\t", $num250, "\t", $num250/$size*100;
print OutFileDist "\n251-300:\t", $num300, "\t", $num300/$size*100;
print OutFileDist "\n300-350:\t", $num350, "\t", $num350/$size*100;
print OutFileDist "\n351-400:\t", $num400, "\t", $num400/$size*100;
print OutFileDist "\n401\> :\t", $num401, "\t", $num401/$size*100, "\n";

close InFile;
close OutFileDist;

## motif search
## Note: Z means stop
%base2aa = ("AAA" => "K", "AAC" => "N", "AAG" => "K", "AAT" => "N", "ACA"
=> "T", "ACC" => "T", "ACG" => "T", "ACT" => "T", "AGA" => "R", "AGC" => "S",
"AGG" => "R", "AGT" => "S", "ATA" => "I", "ATC" => "I", "ATG" => "M", "ATT"
=> "I", "CAA" => "Q", "CAC" => "H", "CAG" => "Q", "CAT" => "H", "CCA" => "P",
"CCC" => "P", "CCG" => "P", "CCT" => "P", "CGA" => "R", "CGC" => "R", "CGG"
=> "R", "CGT" => "R", "CTA" => "L", "CTC" => "L", "CTG" => "L", "CTT" => "L",
"GAA" => "E", "GAC" => "D", "GAG" => "E", "GAT" => "D", "GCA" => "A", "GCC"
=> "A", "GCG" => "A", "GCT" => "A", "GGA" => "G", "GGC" => "G", "GGG" =>
"G", "GGT" => "G", "GTA" => "V", "GTC" => "V", "GTG" => "V", "GTT" => "V",
"TAA" => "Z", "TAC" => "Y", "TAG" => "Z", "TAT" => "Y", "TCA" => "S", "TCC"
=> "S", "TCG" => "S", "TCT" => "S", "TGA" => "Z", "TGC" => "C", "TGG" => "W",
"TGT" => "C", "TTA" => "L", "TTC" => "F", "TTG" => "L", "TTT" => "F");
```

```
open OutFile_CDRH3, ">$infile"._CDR3H.txt' or die $!;
open OutFile_VHFNA, ">$infile-VH.fna" or die $!;
open OutFile_CDRL3, ">$infile"._CDR3L.txt' or die $!;
open OutFile_VLFNA, ">$infile-VL.fna" or die $!;

for ($i=0 ; $i<$size; $i++) {
 $current = $SEQ[$i]; chomp $current;
 $current_r = $current;
 $current_r =~ tr/ACGT[]/TGCA][/; $current_r = reverse($current_r);
 $len=length($current);

 @cdna[1]=""; @cdna[2]=""; @cdna[3]=""; @cdna[4]=""; @cdna[5]=""; @cdna[6]="";
 @AA[1]=""; $cdna[1] = substr ($current, 0); for ($j=0; $j<$len; $j=$j+3) {$triplet =
substr ($current, $j, 3); if ($triplet =~ "N") {$aa="X";} else {$aa=$base2aa{$triplet};}
@AA[1]=@AA[1].$aa;}
 @AA[2]=""; $cdna[2] = substr ($current, 1); for ($j=1; $j<$len; $j=$j+3) {$triplet =
substr ($current, $j, 3); if ($triplet =~ "N") {$aa="X";} else {$aa=$base2aa{$triplet};}
@AA[2]=@AA[2].$aa;}
 @AA[3]=""; $cdna[3] = substr ($current, 2); for ($j=2; $j<$len; $j=$j+3) {$triplet =
substr ($current, $j, 3); if ($triplet =~ "N") {$aa="X";} else {$aa=$base2aa{$triplet};}
@AA[3]=@AA[3].$aa;}
 @AA[4]=""; $cdna[4] = substr ($current_r, 0); for ($j=0; $j<$len; $j=$j+3) {$triplet =
substr ($current_r, $j, 3); if ($triplet =~ "N") {$aa="X";} else {$aa=$base2aa{$triplet};}
@AA[4]=@AA[4].$aa;}
 @AA[5]=""; $cdna[5] = substr ($current_r, 1); for ($j=1; $j<$len; $j=$j+3) {$triplet =
substr ($current_r, $j, 3); if ($triplet =~ "N") {$aa="X";} else {$aa=$base2aa{$triplet};}
@AA[5]=@AA[5].$aa;}
 @AA[6]=""; $cdna[6] = substr ($current_r, 2); for ($j=2; $j<$len; $j=$j+3) {$triplet =
substr ($current_r, $j, 3); if ($triplet =~ "N") {$aa="X";} else {$aa=$base2aa{$triplet};}
@AA[6]=@AA[6].$aa;}

## setting the motif for the motif probability
 $max_P_H1=1e-99; $max_P_L1=1e-99; $max_P_H2=1e-99; $max_P_L2=1e-99;
 for ($j=1; $j<=6; $j++)
 {$current_AA=@AA[$j]; $len_AA=length ($current_AA);
if ($len_AA <10)
{$best_segment_H1="BLANK";$best_segment_L1="BLANK";$best_segment_H2="BL
ANK";$best_segment_L2="BLANK";}
else     {for ($k=0; $k<($len_AA-10); $k++)
          {$segment = substr ($current_AA, $k, 10); $p_H=1e-99; $p_L=1e-99;
```

```
                for ($kk=0; $kk<10; $kk++) {$m[$kk+1] = substr ($segment,$kk,1);}

                if ($m[1] eq "E") {$p_H=0.85;} elsif ($m[1] eq "D") {$p_H=0.08;} else
{$p_H=0.07/18;}
                if ($m[1] eq "E") {$p_L=0.90;} elsif ($m[1] eq "D") {$p_L=0.05;} else
{$p_L=0.05/18;}

                if ($m[2] eq "D") {$p_H=$p_H*0.98;} else {$p_H=$p_H*0.02/19;}
                if ($m[2] eq "D") {$p_L=$p_L*0.98;} else {$p_L=$p_L*0.02/19;}

                if ($m[6] eq "Y") {$p_H=$p_H*0.97;} elsif ($m[6] eq "F")
{$p_H=$p_H*0.01;} else {$p_H=$p_H*0.02/18;}
                if ($m[6] eq "Y") {$p_L=$p_L*0.95;} else {$p_L=$p_L*0.05/19;}

                if ($m[7] eq "Y") {$p_H=$p_H*0.80;} elsif ($m[7] eq "F")
{$p_H=$p_H*0.18;} else {$p_H=$p_H*0.02/18;}
                if ($m[7] eq "Y") {$p_L=$p_L*0.70;} elsif ($m[7] eq "F")
{$p_L=$p_L*0.25;} else {$p_L=$p_L*0.05/18;}

                if ($m[8] eq "C") {$p_H=$p_H*0.99;} else {$p_H=$p_H*0.01/19;}
                if ($m[8] eq "C") {$p_L=$p_L*0.98;} else {$p_L=$p_L*0.02/19;}

                if ($m[9] eq "A") {$p_H=$p_H*0.80;} elsif ($m[9] eq "T")
{$p_H=$p_H*0.08;} elsif ($m[9] eq "V") {$p_H=$p_H*0.04;} else
{$p_H=$p_H*0.08/17;}
                if ($m[9] eq "Q") {$p_L=$p_L*0.50;} elsif ($m[9] eq "H")
{$p_L=$p_L*0.10;} elsif (($m[9] eq "A") or ($m[9] eq "L") or ($m[9] eq "F") or ($m[9]
eq "S")) {$p_L=$p_L*0.05;} else {$p_L=$p_L*0.20/14;}

                if ($m[10] eq "R") {$p_H=$p_H*0.80;} else {$p_H=$p_H*0.20/19;}
                if ($m[10] eq "Q") {$p_L=$p_L*0.80;} elsif ($m[10] eq "H")
{$p_L=$p_L*0.07;} elsif ($m[10] eq "L") {$p_L=$p_L*0.04;} else
{$p_L=$p_L*0.09/17;}


                if ($p_H>$max_P_H1) {$max_P_H1=$p_H;
$best_segment_H1=$segment; $loc1_H1=$j; $loc2_H1=$k; }
                if ($p_L>$max_P_L1) {$max_P_L1=$p_L;
$best_segment_L1=$segment; $loc1_L1=$j; $loc2_L1=$k; }
        }

## motifs search
```

```
for ($k=0; $k<($len_AA-6); $k++)
        {$segment = substr ($current_AA, $k, 6); $p_H=1e-99; $p_L=1e-99;

                for ($kk=0; $kk<6; $kk++) {$m[$kk+1] = substr ($segment,$kk,1);}

                if ($m[1] eq "Y") {$p_H=0.75;} elsif ($m[1] eq "V") {$p_H=0.10;} else
{$p_H=0.15/18;}
                if ($m[1] eq "T") {$p_L=0.92;} elsif ($m[1] eq "V") {$p_L=0.05;} else
{$p_L=0.03/18;}

                if ($m[2] eq "W") {$p_H=$p_H*0.99;} else {$p_H=$p_H*0.01/19;}
                if ($m[2] eq "F") {$p_L=$p_L*0.99;} else {$p_L=$p_L*0.01/19;}

                if ($m[3] eq "G") {$p_H=$p_H*0.98;$p_L=$p_L*0.99;} else
{$p_H=$p_H*0.02/19;$p_L=$p_L*0.01/19;}

                if ($m[4] eq "Q") {$p_H=$p_H*0.80;} elsif ($m[4] eq "A")
{$p_H=$p_H*0.08;} elsif ($m[4] eq "T") {$p_H=$p_H*0.04;} else
{$p_H=$p_H*0.08/17;}
                if ($m[4] eq "G") {$p_L=$p_L*0.55;} elsif ($m[4] eq "A")
{$p_L=$p_L*0.25;} elsif ($m[4] eq "S") {$p_L=$p_L*0.10;} else
{$p_L=$p_L*0.10/17;}

                if ($m[5] eq "G") {$p_H=$p_H*0.99;$p_L=$p_L*0.98;} else
{$p_H=$p_H*0.01/19;$p_L=$p_L*0.02/19;}

                if ($m[6] eq "T") {$p_H=$p_H*0.98; $p_L=$p_L*0.98;} else
{$p_H=$p_H*0.02/19;$p_L=$p_L*0.02/19;}

                if ($p_H>$max_P_H2) {$max_P_H2=$p_H;
$best_segment_H2=$segment; $loc1_H2=$j; $loc2_H2=$k; }
                if ($p_L>$max_P_L2) {$max_P_L2=$p_L;
$best_segment_L2=$segment; $loc1_L2=$j; $loc2_L2=$k; }
         }
        }
 }


$flag=0;
if ((($max_P_L1*$max_P_L2>1e-10) or ($max_P_L1>1e-3) or (($max_P_L1>1e-7) and
($max_P_L2>1e-5))) and ($max_P_H2<1e-6) and ($loc1_L1 == $loc1_L2) and
($loc2_L1<$loc2_L2))
```

```
        {print OutFile_CDRL3 substr(@AA[$loc1_L1], $loc2_L1+8, $loc2_L2-
$loc2_L1-8+1), "\t", @cdna[$loc1_L1], "\n";
        #print OutFile_L $best_segment_L1, "\t", $best_segment_L2, "\t",
@AA[$loc1_L1], "\n";
        print $i, "\t\t", $best_segment_L1, "\t", $best_segment_L2, "\n";
        $flag++;
        print OutFile_VLFNA "$label[$i]$SEQ[$i]\n";}
elsif ((($max_P_H1*$max_P_H2>1e-11) or ($max_P_H1>1e-4) or (($max_P_H1>1e-7)
and ($max_P_H2>1e-5))) and ($max_P_L2<1e-5) and ($loc1_H1 == $loc1_H2) and
($loc2_H1<$loc2_H2))
        {print OutFile_CDRH3 substr(@AA[$loc1_H1], $loc2_H1+8, $loc2_H2-
$loc2_H1-8+1), "\t", @cdna[$loc1_H1], "\n";
        #print OutFile_H $best_segment_H1, "\t", $best_segment_H2, "\t",
@AA[$loc1_H1], "\n";
        print $i, "\t", $best_segment_H1, "\t", $best_segment_H2, "\n";
        $flag++;
        print OutFile_VHFNA "$label[$i]$SEQ[$i]\n";}
#if ($flag==0) {print OutFile_Junk $current, "\n";}
#if ($flag==2) {print OutFile_Double $current, "\n"; print "-----Double \n";}

 }

close OutFile_CDRH3;
close OutFile_CDRL3;
close OutFile_VLFNA;
close OutFile_VHFNA;


## Unique count of the CDR3s and output sorted results
@CDR3L=();
@CDR3H=();
$infile3L="$infile"._CDR3L.txt';
$infile3H="$infile"._CDR3H.txt';
open(InFile3L, $infile3L) or die $!;
open(InFile3H, $infile3H) or die $!;
open OutFile3L, ">$infile"._CDR3L_UNIQUE.txt' or die $!;
open OutFile3H, ">$infile"._CDR3H_UNIQUE.txt' or die $!;

while (my $line=<InFile3L>)
{
chomp($line);
if ($line =~ m/^(.*)\t(.*)/) {push(@CDR3L,$1);}
```

```
}

while (my $line=<InFile3H>)
{
chomp($line);
if ($line =~ m/^(.*)\t(.*)/) {push(@CDR3H,$1);}
}

my %count3L;
map {$count3L{$_}++} @CDR3L;
map {print OutFile3L "$_ \t $count3L{$_}\t",$count3L{$_}/@CDR3L,"\n"} sort keys
(%count3L);

my %count3H;
map {$count3H{$_}++} @CDR3H;
map {print OutFile3H "$_ \t $count3H{$_}\t",$count3H{$_}/@CDR3H,"\n"} sort keys
(%count3H);

close OutFile3L;
close OutFile3H;

$uniqueLfile="$infile"._CDR3L_UNIQUE.txt';
@a=`sort -t $\'\t\' -k2 -nr $uniqueLfile`;
open OutFile3L, ">$infile"._CDR3L_UNIQUE.txt' or die $!;
print OutFile3L @a;
$uniqueHfile="$infile"._CDR3H_UNIQUE.txt';
@b=`sort -t $\'\t\' -k2 -nr $uniqueHfile`;
open OutFile3H, ">$infile"._CDR3H_UNIQUE.txt' or die $!;
print OutFile3H @b;

close InFile3L;
close InFile3H;
close OutFile3L;
close OutFile3H;


## Generate FASTA files for the top specified ranked sequences on VL and VH
## These files can be used for multiple sequence alignments
## The consensus sequences can be used for synthetic genes and candidate selection
@topL=();
@topH=();
@CDR3L=();
```

```
@CDR3H=();
@CDNA3L=();
@CDNA3H=();

open(InFile3L1, "$infile"._CDR3L_UNIQUE.txt') or die $!;
open(InFile3L2, "$infile"._CDR3L.txt') or die $!;
open(InFile3H1, "$infile"._CDR3H_UNIQUE.txt') or die $!;
open(InFile3H2, "$infile"._CDR3H.txt') or die $!;

for (my $i=0; $i<$topnum; $i++)
{
 my $line1=<InFile3L1>;
 chomp($line1);
 if ($line1 =~ m/^(.*)\s\t/) {push(@topL,$1);}
}

while (my $line2=<InFile3L2>)
{
 chomp($line2);
 if ($line2 =~ m/^(.*)\t(.*)/) {push(@CDR3L,$1); push(@CDNA3L,$2);}
}

my $size=@CDR3L;
for (my $i=0; $i<$topnum; $i++)
{
my $n=$i+1;
open OutFileL, ">$n-$infile-VL-$topL[$i].fna" or die $!;
my $num=0;
for (my $j=0; $j<=$size; $j++)
{
if ($CDR3L[$j] eq $topL[$i]) {$num++; print OutFileL "\>Sequence
$num\n$CDNA3L[$j]\n"};
print $j,"running...\n";
}
close OutFileL;
}


for (my $i=0; $i<$topnum; $i++)
{
 my $line1=<InFile3H1>;
 chomp($line1);
```

```
 if ($line1 =~ m/^(.*)\s\t/) {push(@topH,$1);}
}

while (my $line2=<InFile3H2>)
{
 chomp($line2);
 if ($line2 =~ m/^(.*)\t(.*)/) {push(@CDR3H,$1); push(@CDNA3H,$2);}
}

my $size=@CDR3H;
for (my $i=0; $i<$topnum; $i++)
{
my $n=$i+1;
open OutFileH, ">$n-$infile-VH-$topH[$i].fna" or die $!;
my $num=0;
for (my $j=0; $j<=$size; $j++)
{
if ($CDR3H[$j] eq $topH[$i]) {$num++; print OutFileH "\>Sequence
$num\n$CDNA3H[$j]\n"};
print $j,"running...\n";
}
close OutFileH;
}

close InFile3L1;
close InFile3L2;
close InFile3H1;
close InFile3H2;
```

GENE CONVERSION (PERL)

```
#! /usr/bin/perl
# This is a script for identifying the best stretch of a query sequence that is of gene conversion in rabbit (comparing sequences between IGHV1S40 vs other GC reference donors)
# Note: please update the reference name to match your IgBlast reference gene name in the results file. i.e. replace IGHV1S40 with any name that is your base reference genes
# IgBlast command example (assuming blastDB for the reference gene made): igblastn -germline_db_V ./rabbit/ref -germline_db_D ./database/human_gl_D -germline_db_J ./database/human_gl_J -query 1S40seqs_untagged.fna -domain_system imgt -outfmt 3 -num_alignments_V 47 > results.txt
# Usage: perl igblast_geneconv.pl igblast-results.txt

use List::MoreUtils 'pairwise';
use List::MoreUtils 'firstidx';
use List::MoreUtils 'lastidx';


$infile=$ARGV[0];
$infile =~ /(.+)\.txt/;
$filename=$1;
open(InFile,$infile) or die "Error opening $infile !\n";
open(OutFile,">$filename-scores.txt") or die $!;
print "What is your based germline name? (i.e. IGHV1S40 for rabbit):";
chomp($refname=<STDIN>);

print OutFile "Query\tReference\tScore\tStart\tEnd\t$refname-gap\tRef-gap\tShort_start\tShort_end\n";


while ($line=<InFile>) {
	chomp($line);

	if ($line =~ /Query= (.+)/) {
		$queryname=$1;
		%ref=();
		%refori=();
		$sectioncnt=0;
		%lastsection=();
		$printchk=0;
		$lastname='';
	}
```

```
	# Sectioning is used for accommodating reference sequence (esp. pseudogenes)
where alignments do not come in until mid-query sequence...
	if ($line =~ /Query_/) {$sectioncnt++;}

	if ($line =~ /^V\s\s/) {
		@tmp=split(/\s/,$line);
		$n=-1;
		foreach $i (@tmp) {
			if ($i ne '') {
				$n++;
				if ($n==3) {$current=$i;}
				if ($n==5) {$currentseq=$i;}
			}
		}

		if ($current eq $lastname) {next;}

		# The section checking is to ensure proper gapping if the reference aligns
at mid-query sequence
		if (($sectioncnt-$lastsection{$current})>1)
{$refori{$current}=$refori{$current}.'-'x(($sectioncnt-$lastsection{$current}-
1)*70).$currentseq;}
		else {$refori{$current}=$refori{$current}.$currentseq;}
		$currentseq =~ tr/\./1/;
		$currentseq =~ tr/[ATGC-]/0/;
		if (($sectioncnt-$lastsection{$current})>1)
{$ref{$current}=$ref{$current}.'0'x(($sectioncnt-$lastsection{$current}-
1)*70).$currentseq;}
		else {$ref{$current}=$ref{$current}.$currentseq;}
		$lastsection{$current}=$sectioncnt;
		$lastname=$current;
	}

	if (($line =~ /^D\s\s/) && ($printchk==0)) {
		@arr2=split(//,$ref{$refname});
		map {
			$key=$_;
			if ($key ne $refname) {
			 @arr1=split(//,$ref{$key});
			 @score=pairwise{$a-$b} @arr1,@arr2;
			 $sum=0;
```

```
                        $startpos=0;
                        $pos=-1;
                        $maxscore=0;
                        $terminatechk=0;
                        foreach $i (@score) {
                                $pos++;
                                $sum=$sum+$i;
                                if ($sum > $maxscore) {$maxscore=$sum;$chk=1;}
                                # $i=1 means matching the non-based ref and $i=-1 means
matching the based ref
                                if ($i == 1) {$terminatechk=0;}
                                if ($i == -1) {
                                        if ($terminatechk==0) {$spcfinalend=$pos-1;}
                                        $terminatechk++;
                                }
                                if (($i == -1 && $sum <=1) || $pos+1==@score) {
                                        $sum=0;
                                        if ($chk==1) {
                                                $finalstart=$startpos;
                                                if ($terminatechk > 1)
{$finalend=$spcfinalend;}
                                                else {$finalend=$pos-1;}
                                                $chk=0;
                                        }
                                        $startpos=$pos+1;
                                }
                        }
                        if ($maxscore > 1) {
                                $base=substr($refori{$refname},$finalstart,$finalend-
$finalstart+1);
                                $winner=substr($refori{$key},$finalstart,$finalend-
$finalstart+1);
                                $basegap=$base=~tr/-/-/;
                                $winnergap=$winner=~tr/-/-/;
                                $shortstart=firstidx {$_ == 1}
@score[$finalstart..$finalend];
                                $shortend=lastidx {$_ == 1}
@score[$finalstart..$finalend];
                                $shortstart=$finalstart+$shortstart;
                                $shortend=$finalstart+$shortend;
```

```
                                        print OutFile
"$queryname\t$key\t$maxscore\t$finalstart\t$finalend\t$basegap\t$winnergap\t$shortstar
t\t$shortend\n";
                                }
                                }
                        } sort keys (%ref);
                        # The printchk here is to prevent multiple output of the results when
multiple D-gene results were present from the igblast results;
                        # When reports started, printchk will be 1 and will stop generating results
                        $printchk=1;
                }
}

close InFile;
close OutFile;
```

GENE CONVERSION PERMUTATION (PYTHON)

```
#!/usr/bin/python
#This is a script that processes the gene conversion scores files and IgBlast results to give p-values to the scores
#This will shuffle the nucleotides of the baseref germline and the Gene_Conversion germline [trimming about 15 nucleotides from both ends; i.e. the primer regions]
#Then uses the scoring system to find a score iteratively with the specified number of repeats
#If the score is greater or equal than the current score, it will be counted towards the p-value for randomly generating a score higher or as high as the one observed
#Note: the start stop position is based on python index at 0
#Usage: python igblast_geneconvPval.py scores.txt igblast_results.txt

import sys
import re
from collections import defaultdict
from random import randrange
import numpy as np

filem=re.match(r"(.*)\.txt",sys.argv[1])
filename=filem.group(1)
baseref=raw_input("What is your base reference gene (case sensitive)?")
#This is how many iterations needed to generate the p-values; it can be changed accordingly based on hardware specs
iterations=1000

input=defaultdict(list)
querylist=[]
with open(sys.argv[1]) as f:
        for line in f:
                if "Query\tReference\t" in line:
                        continue
                line=line.replace("\r\n",'')
                tmp=line.split("\t")
                input[tmp[0]]=tmp
                querylist.append(tmp[0])
f.close()

section=0
fout=open("p-value-%s.txt"%filename,"w")
print >>fout,"Query\tReference\tScore\tStart\tEnd\tIGHV1S40-gap\tRef-gap\tShort_start\tShort_end\tlocal-count\tp-value"
```

```
with open(sys.argv[2]) as f:
	for line in f:
		#Checking to see whether the section should begin or not by using the
section flag
		if section==0:
			m=re.search(r"Query=\s(.+)",line)
			if m and (m.group(1) in querylist):
				query=m.group(1)
				section=1 #Allowing to begin checking for blocks to be
extracted
				proceed=0 #This would be the flag to signal when to start
extracting sequences
				tempmat=defaultdict(str)
				print "Working on %s ..."%query
		if section==1:
			tmp=line.split(' ')
			tmp=[x for x in tmp if x]
			if 'Query_' in tmp[0]:
				proceed=1
			#Check when to start extracting the sequence
			if 'V' in tmp[0] and proceed==1:
				if tmp[3]==baseref:
					tempmat[baseref]+=tmp[5]
				if tmp[3]==input[query][1]:
					tempmat[input[query][1]]+=tmp[5]
			if tmp[0]=='D' or tmp[0]=='J':
				#Start permutation/shuffling analysis
				proceed=0
				#making sure the extracted lengths are the same
				while
len(tempmat[baseref])>len(tempmat[input[query][1]]):
					tempmat[input[query][1]]+='-'
				while
len(tempmat[baseref])<len(tempmat[input[query][1]]):
					tempmat[baseref]+='-'
				#Trimming the sequences by 15 nucleotides on both ends

	tempmat[baseref]=tempmat[baseref][15:len(tempmat[baseref])-15]

	tempmat[input[query][1]]=tempmat[input[query][1]][15:len(tempmat[input[query
][1]])-15]
				#transform sequence into scores
```

```
                tempmat[baseref]=re.sub(r'[ATGC-]','0',tempmat[baseref])
                tempmat[baseref]=re.sub(r'\.','1',tempmat[baseref])
                tempmat[input[query][1]]=re.sub(r'[ATGCN-
]','0',tempmat[input[query][1]])

        tempmat[input[query][1]]=re.sub(r'\.','1',tempmat[input[query][1]])
                scoremat=defaultdict(list)
                scoremat[baseref]=[int(c) for c in tempmat[baseref]]
                scoremat[input[query][1]]=[int(c) for c in
tempmat[input[query][1]]]
                #Generating the position Numpy array
                pos=range(len(scoremat[baseref]))
                pos=np.array(pos)
                count=0
                for i in range(iterations):
                        np.random.shuffle(pos)
                        randmat=defaultdict(list)
                        for p in pos:

        randmat[baseref].append(scoremat[baseref][p])

        randmat[input[query][1]].append(scoremat[input[query][1]][p])
                                a=np.array(randmat[input[query][1]])
                                b=np.array(randmat[baseref])
                                c=a-b
                                sum=0
                                maxscore=0
                                currentpos=-1
                                for s in c:
                                        currentpos+=1
                                        sum+=s
                                        if sum>maxscore:
                                                maxscore=sum
                                        if (s==-1 and sum<=1) or
currentpos+1==len(c):
                                                sum=0
                                #print "%s\t\t\t%d"%(input[query][2],maxscore)
                                if maxscore>=int(input[query][2]):
                                        count+=1
                        for q in range(9):
                                print >>fout,"%s\t"%input[query][q],
                        pval=(count*1.0)/iterations
```

```
                    print >>fout,"%d\t%.6f"%(count,pval)
                    section=0
f.close()
fout.close()
```

SEED-BASED CIRCLE SEQUENCING SAMPLE PROCESSING (PYTHON)

```
#!/usr/bin/python

# Note: when generating the k-mer hash table; the primers were used to split the sequence
into chunks and the chunks were used to build the hash table and no linkage between the
chunks
#
# Note: it generates FASTQ output
#
# Function: This program is to process the raw Illumina MiSeq sequences (R1 & R2)
# using the algorithm similar to the Inchworm Algorithm to re-create the transcript
#sequence based on repeats
#
# Motivation: This is intended to filter out the PCR-mediated recombination problem
#where short reads were generated within a cluster causing Sequencing Adaptor polluting
#the read quality in subsequent repeats
#
# Output: Fastq file with the quality score expanded, sequence improved by repeats and
# removed of adaptor contaminants
#
# Usage: python kmer-adapt-filter-primer.py [-h] [-k] [-q] MiSeq_R1.fastq
#MiSeq_R2.fastq

from argparse import ArgumentParser
from Bio import SeqIO
from Bio.Alphabet import generic_dna
from Bio.Seq import Seq
from collections import defaultdict
from math import log10
import re
import sys

# The following function will process the R1,R2 reads of the sequence
# to generate the combined kmer hashtables and usage list and quality score hashtables
# **** Note: The primers were used to split the sequence into chunks and no linkage
between the chunks
# ****The same number of chunks was recovered from the quality sequence as well
def process_seq(inseq1,inseq2,inqual1,inqual2):
        seqdict=defaultdict(int)
        seqlist=[]
        qualdict=defaultdict(list)
```

```
        # Splitting the sequence into chunks; using the CCS 3' primer to uncouple the
3'end-5'end linkage
        inseq1chunks=re.split(r"GTCTCCTGTGAGAATTCCCCGTT",inseq1)
        if len(inseq1chunks)==1:
                inseq1chunks=re.split(r"AACGGGGAATTCTCACAGGAGAC",inseq1)
        # Generate list equivalent of inseq1chunks for the quality scores
        # Remove leadning or trailing primer exact match
        if inseq1chunks[0]=='':
                inqual1=inqual1[23:]
        if inseq1chunks[-1]=='':
                inqual1=inqual1[:-23]
        inseq1chunks=[i for i in inseq1chunks if i != '']
        # Start extracting the quality list
        inqual1chunks=[]
        for i in inseq1chunks:
                inqual1chunks.append(inqual1[0:len(i)])
                if len(i)+23<=len(inqual1):
                        inqual1=inqual1[len(i)+23:]
        for n in range(len(inseq1chunks)):
                inseq1=inseq1chunks[n]
                inqual1=inqual1chunks[n]
                while inseq1:
                        if len(inseq1)>=ksize:
                                tempqual=inqual1[0:ksize]
                                # Checking to see if 80% of bases in the kmer is worse than
the specified error rates
                                # If it is the case, simply bypass that kmer and not store in
the hashtables
                                #
                                if sum([1 if i > pcutoffs else 0 for i in
tempqual])>int(ksize*0.80):
                                        inseq1=inseq1[1:]
                                        inqual1=inqual1[1:]
                                else:
                                        seqdict[inseq1[0:ksize]]+=1
                                        if inseq1[0:ksize] in qualdict:
                                                qualdict[inseq1[0:ksize]]=[a*b for a,b in
zip(qualdict[inseq1[0:ksize]],tempqual)]
                                        else:
                                                qualdict[inseq1[0:ksize]]=tempqual
                                        inseq1=inseq1[1:]
                                        inqual1=inqual1[1:]
```

```
                else:
                        break

        # Splitting the sequence into chunks; using the CCS 3' primer to uncouple the
3'end-5'end linkage
        inseq2chunks=re.split(r"GTCTCCTGTGAGAATTCCCCGTT",inseq2)
        if len(inseq2chunks)==1:
                inseq2chunks=re.split(r"AACGGGGAATTCTCACAGGAGAC",inseq2)
        # Generate list equivalent of inseq1chunks for the quality scores
        # Remove leadning or trailing primer exact match
        if inseq2chunks[0]=='':
                inqual2=inqual2[23:]
        if inseq2chunks[-1]=='':
                inqual2=inqual2[:-23]
        inseq2chunks=[i for i in inseq2chunks if i != '']
        # Start extracting the quality list
        inqual2chunks=[]
        for i in inseq2chunks:
                inqual2chunks.append(inqual2[0:len(i)])
                if len(i)+23<=len(inqual2):
                        inqual2=inqual2[len(i)+23:]
        for n in range(len(inseq2chunks)):
                inseq2=inseq2chunks[n]
                inqual2=inqual2chunks[n]
                while inseq2:
                        if len(inseq2)>=ksize:
                                tempqual=inqual2[0:ksize]
                                if sum([1 if i > pcutoffs else 0 for i in
tempqual])>int(ksize*0.8):
                                        inseq2=inseq2[1:]
                                        inqual2=inqual2[1:]
                                else:
                                        seqdict[inseq2[0:ksize]]+=1
                                        if inseq2[0:ksize] in qualdict:
                                                qualdict[inseq2[0:ksize]]=[a*b for a,b in
zip(qualdict[inseq2[0:ksize]],tempqual)]
                                        else:
                                                qualdict[inseq2[0:ksize]]=tempqual
                                        inseq2=inseq2[1:]
                                        inqual2=inqual2[1:]
                        else:
                                for s in sorted(seqdict,key=seqdict.get,reverse=True):
```

```
                seqlist.append(s)
        return seqdict,seqlist,qualdict


# The following function will grow the contig/transcript using seqdict,seqlist,qualdict
# It will call the branch_seq subroutine if tie occurs to branch out the comparison
def extend_seq(seqdict,seqlist,qualdict):
        nuclist=['A','T','G','C']
        usedlist=[seqlist[0]]
        finalseq=seqlist[0]
        qual=qualdict[seqlist[0]]
        chk=1
        # Extend right
        while chk==1:
                countlist=[0,0,0,0]
                for n in range(4):
                        if finalseq[-(ksize-1):]+nuclist[n] in seqlist and finalseq[-(ksize-
1):]+nuclist[n] not in usedlist:
                                countlist[n]=seqdict[finalseq[-(ksize-1):]+nuclist[n]]
                                usedlist.append(finalseq[-(ksize-1):]+nuclist[n])
                if sum(countlist)==0:
                        chk=0
                elif countlist.count(max(countlist))>1:
                        maxlist=[]
                        for n in range(4):
                                if countlist[n]==max(countlist):
                                        maxlist.append(nuclist[n])

        finalseq,usedlist,qual=branch_seq(finalseq,qual,maxlist,0,seqlist,usedlist,qualdict,
ksize)
                else:
                        qual=qual+[qualdict[finalseq[-(ksize-
1):]+nuclist[countlist.index(max(countlist))]][-1]]
                        finalseq=finalseq+nuclist[countlist.index(max(countlist))]


        chk=1
        # Extend left
        while chk==1:
                countlist=[0,0,0,0]
                for n in range(4):
```

```
                        if nuclist[n]+finalseq[0:(ksize-1)] in seqlist and
nuclist[n]+finalseq[0:(ksize-1)] not in usedlist:
                                countlist[n]=seqdict[nuclist[n]+finalseq[0:(ksize-1)]]
                                usedlist.append(nuclist[n]+finalseq[0:(ksize-1)])
                if sum(countlist)==0:
                        chk=0
                elif countlist.count(max(countlist))>1:
                        maxlist=[]
                        for n in range(4):
                                if countlist[n]==max(countlist):
                                        maxlist.append(nuclist[n])

        finalseq,usedlist,qual=branch_seq(finalseq,qual,maxlist,1,seqlist,usedlist,qualdict,
ksize)
                else:

        qual=[qualdict[nuclist[countlist.index(max(countlist))]+finalseq[0:(ksize-
1)]][0]]+qual
                        finalseq=nuclist[countlist.index(max(countlist))]+finalseq
        return finalseq,qual


# The following function will perform the branch off comparison and return final results
until cumulative max is found
# Or, no absolute max is found so contigs growth is terminated [Conservative measure]
# Note: extdirection is the extension direction <-> to right if extdirection=0 and to left if
extdireciton=1
def branch_seq(seed,seedqual,maxlist,extdirection,seqlist,usedlist,qualdict,ksize):
        nuclist=['A','T','G','C']
        branchlist=[]
        branchcount=[]
        if extdirection==0:
                while True:
                        # Build a list of 1 nt extended k-mer from the branches
                        for i in maxlist:
                                for j in range(4):
                                        branchlist.append(i+nuclist[j])
                                        branchcount.append(0)
                        for n in range(len(branchlist)):
                                if seed[-(ksize-len(branchlist[n])):]+branchlist[n] in seqlist
and seed[-(ksize-len(branchlist[n])):]+branchlist[n] not in usedlist:
```

```
                                        branchcount[n]=seqdict[seed[-(ksize-
len(branchlist[n])):]+branchlist[n]]
                                        usedlist.append(seed[-(ksize-
len(branchlist[n])):]+branchlist[n])
                        # If no matching k-mer found, report previously found seed as
finalseq
                        if sum(branchcount)==0:
                                finalseq=seed
                                qual=seedqual
                                return finalseq,usedlist,qual
                        elif branchcount.count(max(branchcount))>1:
                                passmaxlist=[]
                                for n in range(len(branchcount)):
                                        if branchcount[n]==max(branchcount):
                                                passmaxlist.append(branchlist[n])
                                # Loop back with updated maxlist and reset the branches
variables
                                maxlist=passmaxlist
                                branchlist=[]
                                branchcount=[]
                        else:
                                # Report values when absolute max is found

        lenextend=len(branchlist[branchcount.index(max(branchcount))])
                                # Note the quality of the extension should be a list already;
therefore no need to convert it to list as in the single extension case
                                qual=seedqual+qualdict[seed[-(ksize-
lenextend):]+branchlist[branchcount.index(max(branchcount))]][-lenextend:]

        finalseq=seed+branchlist[branchcount.index(max(branchcount))]
                                return finalseq,usedlist,qual

        if extdirection==1:
                while True:
                        for i in maxlist:
                                for j in range(4):
                                        branchlist.append(nuclist[j]+i)
                                        branchcount.append(0)
                        for n in range(len(branchlist)):
                                if branchlist[n]+seed[0:(ksize-len(branchlist[n]))] in seqlist
and branchlist[n]+seed[0:(ksize-len(branchlist[n]))] not in usedlist:
```

```
        branchcount[n]=seqdict[branchlist[n]+seed[0:(ksize-len(branchlist[n]))]]
                                usedlist.append(branchlist[n]+seed[0:(ksize-
len(branchlist[n]))])
                        if sum(branchcount)==0:
                                finalseq=seed
                                qual=seedqual
                                return finalseq,usedlist,qual
                        elif branchcount.count(max(branchcount))>1:
                                passmaxlist=[]
                                for n in range(len(branchcount)):
                                        if branchcount[n]==max(branchcount):
                                                passmaxlist.append(branchlist[n])
                                maxlist=passmaxlist
                                branchlist=[]
                                branchcount=[]
                        else:

        lenextend=len(branchlist[branchcount.index(max(branchcount))])

        qual=qualdict[branchlist[branchcount.index(max(branchcount))]+seed[0:(ksize-
lenextend)]][0:lenextend]+seedqual

        finalseq=branchlist[branchcount.index(max(branchcount))]+seed
                                return finalseq,usedlist,qual



##############################
# Main function starts here #
##############################

# Arguments parsing
parser=ArgumentParser()
parser.add_argument("-k","--kmersize",type=int,default=11,help='k-mer size (Default:
11; Recommend odd kmer)')
parser.add_argument("-q","--qscoremin",type=int,default=1,help='Quality score minimal
for 4/5 of bases to be considered a k-mer (Default: 1; Essentially imposing no
thredshold)')
parser.add_argument("R1",help='Illumina R1 fastq reads (.fastq)')
parser.add_argument("R2",help='Illumina R2 fastq reads (.fastq)')
args=parser.parse_args()
```

```
# File handles and parameter assignment
R1file=open(args.R1,"rU")
R1parse=SeqIO.parse(R1file,"fastq")
R2file=open(args.R2,"rU")
R2parse=SeqIO.parse(R2file,"fastq")
ksize=args.kmersize
pcutoffs=10**-(args.qscoremin*1.0/10)

# Output file assignment
mtch=re.search(r'(.*).fastq',args.R1)
filename=mtch.group(1)
fout=open("%s-adaptfiltr-k%dq%d.fastq"%(filename,ksize,args.qscoremin),"w")

# Going through the files sequence-by-sequence to use the process_seq and extend_seq
functions
totlen=0
seqcount=0
unbuiltseq=0
for record1 in R1parse:
        record2=R2parse.next()
        seqcount+=1
        seqheader=record1.name
        r1seq=str(record1.seq)
        r2seq=str(Seq(str(record2.seq),generic_dna).reverse_complement())
        r1qual=record1.letter_annotations['phred_quality']
        r2qual=record2.letter_annotations['phred_quality']
        r2qual=r2qual[::-1] # reversing the order of quality score to match the reverse
complemented sequence
        # convert the phred score to probability
        r1qual=[10**-(i*1.0/10) for i in r1qual]
        r2qual=[10**-(i*1.0/10) for i in r2qual]
        # Filtering out short reads
        if len(r1seq)<4*ksize or len(r2seq)<4*ksize:
                continue
        seqdict,seqlist,qualdict=process_seq(r1seq,r2seq,r1qual,r2qual)
        # Provide proper error message if quality score was set too high resulting in no
qualified kmer list
        if len(seqlist)==0:
                #sys.exit("ERROR: The minimal quality score is too strict! No kmer is
qualifiied.")
                unbuiltseq+=1
```

```
                continue
        finalseq,qual=extend_seq(seqdict,seqlist,qualdict)
        # Convert P to quality score then to ASCII code Illumina V1.8 (Ascii based 33)
        # Limiting quality score to max out at 93 ASCII code is '~'
        qual=[chr(int(-10*log10(i)+33)) if i>5.1e-10 else '~' for i in qual]
        qual=''.join(qual)
        print >>fout,"@%s\n%s\n+\n%s"%(seqheader,finalseq,qual)
        totlen+=len(finalseq)

# Reporting some basic info
fout2=open("%s-k%dq%d-log.txt"%(filename,ksize,args.qscoremin),"w")
print >>fout2,"Total number of sequences analyzed: %d"%seqcount
print >>fout2,"Average length: %.4f"%((totlen*1.0)/seqcount)
print >>fout2,"Number of sequences discarded due to strict requirements:
%d"%unbuiltseq

# Closing file handles
R1file.close()
R2file.close()
fout.close()
fout2.close()
```

LIST OF ADDITIONAL SCRIPTS

Due to the limited space in this dissertation, the list below describes some more scripts created for most projects throughout author's graduate program. Please send request to author for these scripts. Note: this is not a complete list and please contact author as other practical scripts might have been created that can serve your needs in handling NGS data or visualization purposes. Also, some scripts were deposited on Appsoma.com under the tag Immunogrep.

Data I/O /pre-processing/data conversion:

- Split FASTA: splitting FASTA file into batches of different number of sequences

- Split FASTQ: splitting FASTQ file into batches of different number of sequences

- FASTQ to FASTA: converting FASTQ format into FASTA format

- GenBank record fetcher: automated script that fetches GenBank records and sequences

Data transformation/extraction:

- IMGT parsing scripts: various versions that transform IMGT outputs into unique antibody output

- Clonotype: combining Usearch program and IMGT V gene assignment for clonotyping antibody sequences

- Protease related scripts: analyze enrichment from protease data

- Isotyping script: analyzing antibody sequences to extract and categorize isotypes based on known isotype constant region sequences

Data visualization:

- Variability plot: tool to process sequencing data and plot the Kabat-Wu variability plot

- Hydropathy plot: tool to process sequencing data and plot the Eisenberg scale hydropathy plot

- Species Richness plot: tool referencing the diversity information to plot the rarefaction curve

References

[1] K. P. Murphy and C. Janeway, *Janeway's Immunobiology*, 7th ed. New York: Garland Science, 2008.

[2] A. M. Silverstein, *A History of Immunology*. Academic Press, 2009.

[3] S. Aggarwal, "What's fueling the biotech engine--2008," *Nat. Biotechnol.*, vol. 27, no. 11, pp. 987–993, Nov. 2009.

[4] S. (Rob) Aggarwal, "What's fueling the biotech engine—2012 to 2013," *Nat. Biotechnol.*, vol. 32, no. 1, pp. 32–39, Jan. 2014.

[5] P. M. LoRusso, D. Weiss, E. Guardino, S. Girish, and M. X. Sliwkowski, "Trastuzumab emtansine: a unique antibody-drug conjugate in development for human epidermal growth factor receptor 2-positive cancer," *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 17, no. 20, pp. 6437–6447, Oct. 2011.

[6] C. Rioufol and G. Salles, "Obinutuzumab for chronic lymphocytic leukemia," *Expert Rev. Hematol.*, vol. 7, no. 5, pp. 533–543, Oct. 2014.

[7] P. M. Alzari, M. B. Lascombe, and R. J. Poljak, "Three-dimensional structure of antibodies," *Annu. Rev. Immunol.*, vol. 6, pp. 555–580, 1988.

[8] P. Eduardo A., "Anatomy of the antibody molecule," *Mol. Immunol.*, vol. 31, no. 3, pp. 169–217, Feb. 1994.

[9] M. J. Niles, L. Matsuuchi, and M. E. Koshland, "Polymer IgM assembly and secretion in lymphoid and nonlymphoid cell lines: evidence that J chain is required for pentamer IgM synthesis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 7, pp. 2884–2888, Mar. 1995.

[10] B. A. Hendrickson, D. A. Conner, D. J. Ladd, D. Kendall, J. E. Casanova, B. Corthesy, E. E. Max, M. R. Neutra, C. E. Seidman, and J. G. Seidman, "Altered hepatic transport of immunoglobulin A in mice lacking the J chain," *J. Exp. Med.*, vol. 182, no. 6, pp. 1905–1911, Dec. 1995.

[11] A. Pincetic, S. Bournazos, D. J. DiLillo, J. Maamary, T. T. Wang, R. Dahan, B.-M. Fiebiger, and J. V. Ravetch, "Type I and type II Fc receptors regulate innate and adaptive immunity," *Nat. Immunol.*, vol. 15, no. 8, pp. 707–716, Aug. 2014.

[12] F. E. van de Geijn, M. Wuhrer, M. H. Selman, S. P. Willemsen, Y. A. de Man, A. M. Deelder, J. M. Hazes, and R. J. Dolhain, "Immunoglobulin G galactosylation and sialylation are associated with pregnancy-induced improvement of rheumatoid arthritis and the postpartum flare: results from a large prospective cohort study," *Arthritis Res. Ther.*, vol. 11, no. 6, p. R193, 2009.

[13] J. E. T. Narciso, I. D. C. Uy, A. B. Cabang, J. F. C. Chavez, J. L. B. Pablo, G. P. Padilla-Concepcion, and E. A. Padlan, "Analysis of the antibody structure based on high-resolution crystallographic studies," *New Biotechnol.*, vol. 28, no. 5, pp. 435–447, Sep. 2011.

[14] C. Chothia and A. M. Lesk, "Canonical structures for the hypervariable regions of immunoglobulins," *J. Mol. Biol.*, vol. 196, no. 4, pp. 901–917, Aug. 1987.

[15] C. Chothia, A. M. Lesk, A. Tramontano, M. Levitt, S. J. Smith-Gill, G. Air, S. Sheriff, E. A. Padlan, D. Davies, W. R. Tulip, P. M. Colman, S. Spinelli, P. M. Alzari, and R. J. Poljak, "Conformations of immunoglobulin hypervariable regions," *Nature*, vol. 342, no. 6252, pp. 877–883, Dec. 1989.
[16] G. Johnson and T. T. Wu, "Kabat Database and its applications: 30 years after the first variability plot," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 214–218, Jan. 2000.
[17] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Jabado-Michaloud, G. Folch, F. Bellahcene, Y. Wu, E. Gemrot, X. Brochet, J. Lane, L. Regnier, F. Ehrenmann, G. Lefranc, and P. Duroux, "IMGT®, the international ImMunoGeneTics information system®," *Nucleic Acids Res.*, vol. 37, no. suppl 1, pp. D1006–D1012, Jan. 2009.
[18] G. C. Ippolito, K. H. Hoi, S. T. Reddy, S. M. Carroll, X. Ge, T. Rogosch, M. Zemlin, L. D. Shultz, A. D. Ellington, C. L. VanDenBerg, and G. Georgiou, "Antibody Repertoires in Humanized NOD-scid-IL2Rγnull Mice and Human B Cells Reveals Human-Like Diversification and Tolerance Checkpoints in the Mouse," *PLoS ONE*, vol. 7, no. 4, Apr. 2012.
[19] B. North, A. Lehmann, and R. L. Dunbrack Jr, "A New Clustering of Antibody CDR Loop Conformations," *J. Mol. Biol.*, vol. 406, no. 2, pp. 228–256, Feb. 2011.
[20] B. J. DeKosky, G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dörner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G. Willson, A. D. Ellington, and G. Georgiou, "High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire," *Nat. Biotechnol.*, vol. 31, no. 2, pp. 166–169, Feb. 2013.
[21] G. Georgiou, G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake, "The promise and challenge of high-throughput sequencing of the antibody repertoire," *Nat. Biotechnol.*, vol. 32, no. 2, pp. 158–168, Feb. 2014.
[22] T. W. LeBien and T. F. Tedder, "B lymphocytes: how they develop and function," *Blood*, vol. 112, no. 5, pp. 1570 –1580, 2008.
[23] F. W. Alt, T. K. Blackwell, R. A. DePinho, M. G. Reth, and G. D. Yancopoulos, "Regulation of genome rearrangement events during lymphocyte differentiation," *Immunol. Rev.*, vol. 89, pp. 5–30, Feb. 1986.
[24] J. Jacob, G. Kelsoe, K. Rajewsky, and U. Weiss, "Intraclonal generation of antibody mutants in germinal centres," *Nature*, vol. 354, no. 6352, pp. 389–392, Dec. 1991.
[25] C. Berek, A. Berger, and M. Apel, "Maturation of the immune response in germinal centers," *Cell*, vol. 67, no. 6, pp. 1121–1129, Dec. 1991.
[26] M. G. McHeyzer-Williams, M. J. McLean, P. A. Lalor, and G. J. Nossal, "Antigen-driven B cell differentiation in vivo.," *J. Exp. Med.*, vol. 178, no. 1, pp. 295–307, Jul. 1993.
[27] Y. Takahashi, P. R. Dutta, D. M. Cerasoli, and G. Kelsoe, "In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. V. Affinity maturation develops in two stages of clonal selection," *J. Exp. Med.*, vol. 187, no. 6, pp. 885–895, Mar. 1998.

[28] S. Tonegawa, C. Steinberg, S. Dube, and A. Bernardini, "Evidence for Somatic Generation of Antibody Diversity," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 71, no. 10, pp. 4027–4031, Oct. 1974.
[29] S. Tonegawa, "Somatic generation of antibody diversity," *Nature*, vol. 302, no. 5909, pp. 575–581, Apr. 1983.
[30] F. Melchers, "The pre-B-cell receptor: selector of fitting immunoglobulin heavy chains for the B-cell repertoire," *Nat. Rev. Immunol.*, vol. 5, no. 7, pp. 578–584, Jul. 2005.
[31] H. Azulay-Debby and D. Melamed, "B cell receptor editing in tolerance and autoimmunity," *Front. Biosci. J. Virtual Libr.*, vol. 12, pp. 2136–2147, 2007.
[32] T. McCaughtry and K. Hogquist, "Central tolerance: what have we learned from mice?," *Semin. Immunopathol.*, vol. 30, no. 4, pp. 399–409, 2008.
[33] D. Nemazee, "Receptor editing in lymphocyte development and central tolerance," *Nat Rev Immunol*, vol. 6, no. 10, pp. 728–740, Oct. 2006.
[34] L. A. Herzenberg, "B-1 cells: the lineage question revisited," *Immunol. Rev.*, vol. 175, pp. 9–22, Jun. 2000.
[35] J. Quintáns and H. Cosenza, "Antibody response to phosphorylcholine in vitro. II. Analysis of T-dependent and T-independent responses," *Eur. J. Immunol.*, vol. 6, no. 6, pp. 399–405, Jun. 1976.
[36] H. W. Schroeder Jr. and L. Cavacini, "Structure and function of immunoglobulins," *J. Allergy Clin. Immunol.*, vol. 125, no. 2, Supplement 2, pp. S41–S52, Feb. 2010.
[37] C. P. Chappell, K. E. Draves, N. V. Giltiay, and E. A. Clark, "Extrafollicular B cell activation by marginal zone dendritic cells drives T cell-dependent antibody responses," *J. Exp. Med.*, vol. 209, no. 10, pp. 1825–1840, Sep. 2012.
[38] B. Hou, P. Saudan, G. Ott, M. L. Wheeler, M. Ji, L. Kuzmich, L. M. Lee, R. L. Coffman, M. F. Bachmann, and A. L. DeFranco, "Selective utilization of Toll-like receptor and MyD88 signaling in B cells for enhancement of the antiviral germinal center response," *Immunity*, vol. 34, no. 3, pp. 375–384, Mar. 2011.
[39] S. P. Kasturi, I. Skountzou, R. A. Albrecht, D. Koutsonanos, T. Hua, H. I. Nakaya, R. Ravindran, S. Stewart, M. Alam, M. Kwissa, F. Villinger, N. Murthy, J. Steel, J. Jacob, R. J. Hogan, A. García-Sastre, R. Compans, and B. Pulendran, "Programming the magnitude and persistence of antibody responses with innate immunity," *Nature*, vol. 470, no. 7335, pp. 543–547, Feb. 2011.
[40] G. D. Victora and M. C. Nussenzweig, "Germinal centers," *Annu. Rev. Immunol.*, vol. 30, pp. 429–457, 2012.
[41] T. Yoshida, H. Mei, T. Dörner, F. Hiepe, A. Radbruch, S. Fillatreau, and B. F. Hoyer, "Memory B and memory plasma cells," *Immunol. Rev.*, vol. 237, no. 1, pp. 117–139, Aug. 2010.
[42] N. Pelletier and M. G. McHeyzer-Williams, "B cell memory: how to start and when to end," *Nat Immunol*, vol. 10, no. 12, pp. 1233–1235, Dec. 2009.

[43] K. A. Pape, J. J. Taylor, R. W. Maul, P. J. Gearhart, and M. K. Jenkins, "Different B Cell Populations Mediate Early and Late Memory During an Endogenous Immune Response," *Science*, vol. 331, no. 6021, pp. 1203–1207, Mar. 2011.
[44] T. Kaji, A. Ishige, M. Hikida, J. Taka, A. Hijikata, M. Kubo, T. Nagashima, Y. Takahashi, T. Kurosaki, M. Okada, O. Ohara, K. Rajewsky, and T. Takemori, "Distinct cellular pathways select germline-encoded and somatically mutated antibodies into immunological memory," *J. Exp. Med.*, Oct. 2012.
[45] M. McHeyzer-Williams, S. Okitsu, N. Wang, and L. McHeyzer-Williams, "Molecular programming of B cell memory," *Nat Rev Immunol*, vol. 12, no. 1, pp. 24–34, Jan. 2012.
[46] D. Frölich, C. Giesecke, H. E. Mei, K. Reiter, C. Daridon, P. E. Lipsky, and T. Dörner, "Secondary Immunization Generates Clonally Related Antigen-Specific Plasma Cells and Memory B Cells," *J. Immunol.*, vol. 185, no. 5, pp. 3103–3110, Sep. 2010.
[47] K. L. Good-Jacobson and M. J. Shlomchik, "Plasticity and Heterogeneity in the Generation of Memory B Cells and Long-Lived Plasma Cells: The Influence of Germinal Center Interactions and Dynamics," *J. Immunol.*, vol. 185, no. 6, pp. 3117 –3125, 2010.
[48] E. J. Blink, A. Light, A. Kallies, S. L. Nutt, P. D. Hodgkin, and D. M. Tarlinton, "Early appearance of germinal center-derived memory B cells and plasma cells in blood after primary immunization," *J. Exp. Med.*, vol. 201, no. 4, pp. 545–554, Feb. 2005.
[49] H. Toyama, S. Okada, M. Hatano, Y. Takahashi, N. Takeda, H. Ichii, T. Takemori, Y. Kuroda, and T. Tokuhisa, "Memory B cells without somatic hypermutation are generated from Bcl6-deficient B cells," *Immunity*, vol. 17, no. 3, pp. 329–339, Sep. 2002.
[50] A. Radbruch, G. Muehlinghaus, E. O. Luger, A. Inamine, K. G. C. Smith, T. Dorner, and F. Hiepe, "Competence and competition: the challenge of becoming a long-lived plasma cell," *Nat Rev Immunol*, vol. 6, no. 10, pp. 741–750, Oct. 2006.
[51] V. T. Chu, A. Beller, T. T. N. Nguyen, G. Steinhauser, and C. Berek, "The Long-Term Survival of Plasma Cells," *Scand. J. Immunol.*, vol. 73, no. 6, pp. 508–511, Jun. 2011.
[52] M. K. Slifka, R. Antia, J. K. Whitmire, and R. Ahmed, "Humoral Immunity Due to Long-Lived Plasma Cells," *Immunity*, vol. 8, no. 3, pp. 363–372, Mar. 1998.
[53] R. A. Manz, S. Arce, G. Cassese, A. E. Hauser, F. Hiepe, and A. Radbruch, "Humoral immunity and long-lived plasma cells," *Curr. Opin. Immunol.*, vol. 14, no. 4, pp. 517–521, Aug. 2002.
[54] K. A. Fairfax, A. Kallies, S. L. Nutt, and D. M. Tarlinton, "Plasma cell development: from B-cell subsets to long-term survival niches," *Semin. Immunol.*, vol. 20, no. 1, pp. 49–58, Feb. 2008.

[55] K. Moser, K. Tokoyoda, A. Radbruch, I. MacLennan, and R. A. Manz, "Stromal niches, plasma cell differentiation and survival," *Curr. Opin. Immunol.*, vol. 18, no. 3, pp. 265–270, Jun. 2006.
[56] T. Höfer, G. Muehlinghaus, K. Moser, T. Yoshida, H. E. Mei, K. Hebel, A. Hauser, B. Hoyer, E. O. Luger, T. Dörner, R. A. Manz, F. Hiepe, and A. Radbruch, "Adaptation of humoral memory," *Immunol. Rev.*, vol. 211, no. 1, pp. 295–302, 2006.
[57] D. Tarlinton, A. Radbruch, F. Hiepe, and T. Dörner, "Plasma cell differentiation and survival," *Curr. Opin. Immunol.*, vol. 20, no. 2, pp. 162–169, Apr. 2008.
[58] M. Odendahl, H. Mei, B. F. Hoyer, A. M. Jacobi, A. Hansen, G. Muehlinghaus, C. Berek, F. Hiepe, R. Manz, A. Radbruch, and T. Dorner, "Generation of migratory antigen-specific plasma blasts and mobilization of resident plasma cells in a secondary immune response," *Blood*, vol. 105, no. 4, pp. 1614–1621, Feb. 2005.
[59] F. Ho, J. E. Lortan, I. C. MacLennan, and M. Khan, "Distinct short-lived and long-lived antibody-producing cell populations," *Eur. J. Immunol.*, vol. 16, no. 10, pp. 1297–1301, Oct. 1986.
[60] R. A. Manz, A. Thiel, and A. Radbruch, "Lifetime of plasma cells in the bone marrow," *Nature*, vol. 388, no. 6638, pp. 133–134, Jul. 1997.
[61] R. A. Manz, M. Löhning, G. Cassese, A. Thiel, and A. Radbruch, "Survival of long-lived plasma cells is independent of antigen.," *Int. Immunol.*, vol. 10, no. 11, pp. 1703–1711, Nov. 1998.
[62] I. J. Amanna, N. E. Carlson, and M. K. Slifka, "Duration of Humoral Immunity to Common Viral and Vaccine Antigens," *N. Engl. J. Med.*, vol. 357, no. 19, pp. 1903–1915, 2007.
[63] F. Matsuda, K. Ishii, P. Bourvagnet, K. i Kuma, H. Hayashida, T. Miyata, and T. Honjo, "The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus," *J. Exp. Med.*, vol. 188, no. 11, pp. 2151–2162, Dec. 1998.
[64] K. F. Schäble and H. G. Zachau, "The variable genes of the human immunoglobulin kappa locus," *Biol. Chem. Hoppe. Seyler*, vol. 374, no. 11, pp. 1001–1022, Nov. 1993.
[65] H. G. Zachau, "The immunoglobulin kappa locus-or-what has been learned from looking closely at one-tenth of a percent of the human genome," *Gene*, vol. 135, no. 1–2, pp. 167–173, Dec. 1993.
[66] J. P. Frippiat, S. C. Williams, I. M. Tomlinson, G. P. Cook, D. Cherif, D. Le Paslier, J. E. Collins, I. Dunham, G. Winter, and M. P. Lefranc, "Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2," *Hum. Mol. Genet.*, vol. 4, no. 6, pp. 983–991, Jun. 1995.
[67] V. Giudicelli and M.-P. Lefranc, "Ontology for immunogenetics: the IMGT-ONTOLOGY," *Bioinformatics*, vol. 15, no. 12, pp. 1047 –1054, Dec. 1999.
[68] V. Giudicelli, P. Duroux, C. Ginestoux, G. Folch, J. Jabado-Michaloud, D. Chaume, and M.-P. Lefranc, "IMGT/LIGM-DB, the IMGT® comprehensive database of

immunoglobulin and T cell receptor nucleotide sequences," *Nucleic Acids Res.*, vol. 34, no. suppl 1, pp. D781–D784, Jan. 2006.
[69] S. D. Boyd, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. Z. Fire, "Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing," *Sci. Transl. Med.*, vol. 1, no. 12, pp. 12ra23–12ra23, Dec. 2009.
[70] J. Glanville, W. Zhai, J. Berka, D. Telman, G. Huerta, G. R. Mehta, I. Ni, L. Mei, P. D. Sundar, G. M. R. Day, D. Cox, A. Rajpal, and J. Pons, "Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire," *Proc. Natl. Acad. Sci.*, vol. 106, no. 48, pp. 20216 – 20221, Dec. 2009.
[71] R. Arnaout, W. Lee, P. Cahill, T. Honan, T. Sparrow, M. Weiand, C. Nusbaum, K. Rajewsky, and S. B. Koralov, "High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans," *PLoS ONE*, vol. 6, no. 8, p. e22365, Aug. 2011.
[72] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci.*, vol. 74, no. 12, pp. 5463 –5467, Dec. 1977.
[73] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, Oct. 2004.
[74] W. W. Soon, M. Hariharan, and M. P. Snyder, "High-throughput sequencing for biology and medicine," *Mol. Syst. Biol.*, vol. 9, no. 1, Jan. 2013.
[75] G. Alterovitz, R. Benson, and M. F. Ramoni, *Automation in proteomics and genomics: an engineering case-based approach*. John Wiley and Sons, 2009.
[76] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1135–1145, Oct. 2008.
[77] M. Quail, M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC Genomics*, vol. 13, no. 1, p. 341, Jul. 2012.
[78] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén, "Real-time DNA sequencing using detection of pyrophosphate release," *Anal. Biochem.*, vol. 242, no. 1, pp. 84–89, Nov. 1996.
[79] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg,

"Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, Sep. 2005.
[80] T. C. Glenn, "Field guide to next-generation DNA sequencers," *Mol. Ecol. Resour.*, vol. 11, no. 5, pp. 759–769, Sep. 2011.
[81] R. Williams, S. G. Peisajovich, O. J. Miller, S. Magdassi, D. S. Tawfik, and A. D. Griffiths, "Amplification of complex gene libraries by emulsion PCR," *Nat. Methods*, vol. 3, no. 7, pp. 545–550, Jul. 2006.
[82] I. F. Bronner, M. A. Quail, D. J. Turner, and H. Swerdlow, "Improved Protocols for Illumina Sequencing," *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al*, vol. 0 18, Jul. 2009.
[83] S. T. Reddy, X. Ge, A. E. Miklos, R. A. Hughes, S. H. Kang, K. H. Hoi, C. Chrysostomou, S. P. Hunicke-Smith, B. L. Iverson, P. W. Tucker, A. D. Ellington, and G. Georgiou, "Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells," *Nat Biotech*, vol. 28, no. 9, pp. 965–969, 2010.
[84] Y. Wine, D. R. Boutz, J. J. Lavinder, A. E. Miklos, R. A. Hughes, K. H. Hoi, S. T. Jung, A. P. Horton, E. M. Murrin, A. D. Ellington, E. M. Marcotte, and G. Georgiou, "Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response," *Proc. Natl. Acad. Sci.*, vol. 110, no. 8, pp. 2993–2998, Feb. 2013.
[85] J. J. Lavinder, Y. Wine, C. Giesecke, G. C. Ippolito, A. P. Horton, O. I. Lungu, K. H. Hoi, B. J. DeKosky, E. M. Murrin, M. M. Wirth, A. D. Ellington, T. Dörner, E. M. Marcotte, D. R. Boutz, and G. Georgiou, "Identification and characterization of the constituent human serum antibodies elicited by vaccination," *Proc. Natl. Acad. Sci.*, vol. 111, no. 6, pp. 2259–2264, Feb. 2014.
[86] Y.-C. Tan, S. Kongpachith, L. K. Blum, C.-H. Ju, L. J. Lahey, D. R. Lu, X. Cai, C. A. Wagner, T. M. Lindstrom, J. Sokolove, and W. H. Robinson, "Barcode-Enabled Sequencing of Plasmablast Antibody Repertoires in Rheumatoid Arthritis," *Arthritis Rheumatol.*, vol. 66, no. 10, pp. 2706–2715, Oct. 2014.
[87] M. Michaeli, H. Noga, H. Tabibian-Keissar, I. Barshack, and R. Mehr, "Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing," *Front. Immunol.*, vol. 3, 2012.
[88] B. A. Gaëta, H. R. Malming, K. J. L. Jackson, M. E. Bain, P. Wilson, and A. M. Collins, "iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences," *Bioinformatics*, vol. 23, no. 13, pp. 1580–1587, Jul. 2007.
[89] X. Brochet, M.-P. Lefranc, and V. Giudicelli, "IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. W503–W508, Jul. 2008.

[90] X. Wang, D. Wu, S. Zheng, J. Sun, L. Tao, Y. Li, and Z. Cao, "Ab-origin: an enhanced tool to identify the sourcing gene segments in germline for rearranged antibodies," *BMC Bioinformatics*, vol. 9 Suppl 12, p. S20, 2008.
[91] S. Munshaw and T. B. Kepler, "SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements," *Bioinformatics*, vol. 26, no. 7, pp. 867–872, Apr. 2010.
[92] J. Ye, N. Ma, T. L. Madden, and J. M. Ostell, "IgBLAST: an immunoglobulin variable domain sequence analysis tool," *Nucleic Acids Res.*, vol. 41, no. Web Server issue, pp. W34–40, Jul. 2013.
[93] S. D'Angelo, J. Glanville, F. Ferrara, L. Naranjo, C. D. Gleasner, X. Shen, A. R. Bradbury, and C. Kiss, "The antibody mining toolbox," *mAbs*, vol. 6, no. 1, pp. 160–172, Jan. 2014.
[94] R. C. Edgar, "Search and Clustering Orders of Magnitude Faster Than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Oct. 2010.
[95] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010.
[96] Z. Chen, A. M. Collins, Y. Wang, and B. A. Gaëta, "Clustering-based identification of clonally-related immunoglobulin gene sequence sets," *Immunome Res.*, vol. 6, no. Suppl 1, p. S4, Sep. 2010.
[97] M. Michaeli, M. Barak, L. Hazanov, H. Noga, and R. Mehr, "Automated analysis of immunoglobulin genes from high-throughput sequencing: life without a template," *J. Clin. Bioinforma.*, vol. 3, no. 1, p. 15, Aug. 2013.
[98] T. Magoč and S. L. Salzberg, "FLASH: fast length adjustment of short reads to improve genome assemblies," *Bioinformatics*, vol. 27, no. 21, pp. 2957–2963, Nov. 2011.
[99] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, "PEAR: a fast and accurate Illumina Paired-End reAd mergeR," *Bioinformatics*, vol. 30, no. 5, pp. 614–620, Mar. 2014.
[100] J. Benichou, J. Glanville, E. T. L. Prak, R. Azran, T. C. Kuo, J. Pons, C. Desmarais, L. Tsaban, and Y. Louzoun, "The Restricted DH Gene Reading Frame Usage in the Expressed Human Antibody Repertoire Is Selected Based upon its Amino Acid Content," *J. Immunol.*, Apr. 2013.
[101] M. Faham, J. Zheng, M. Moorhead, V. E. H. Carlton, P. Stow, E. Coustan-Smith, C.-H. Pui, and D. Campana, "Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia," *Blood*, vol. 120, no. 26, pp. 5173–5180, Dec. 2012.
[102] D. Wu, A. Sherwood, J. R. Fromm, S. S. Winter, K. P. Dunsmore, M. L. Loh, H. A. Greisman, D. E. Sabath, B. L. Wood, and H. Robins, "High-Throughput Sequencing Detects Minimal Residual Disease in Acute T Lymphoblastic Leukemia," *Sci. Transl. Med.*, vol. 4, no. 134, pp. 134ra63–134ra63, May 2012.

[103] R. Küppers, U. Klein, M. L. Hansmann, and K. Rajewsky, "Cellular origin of human B-cell lymphomas," *N. Engl. J. Med.*, vol. 341, no. 20, pp. 1520–1529, Nov. 1999.
[104] T. Rogosch, S. Kerzel, K. H. Hoi, Z. Zhang, G. C. Ippolito, and M. Zemlin, "Immunoglobulin Analysis Tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts," *Front. B Cell Biol.*, vol. 3, p. 176, 2012.
[105] L. Ohm-Laursen, M. Nielsen, S. R. Larsen, and T. Barington, "No evidence for the use of DIR, D–D fusions, chromosome 15 open reading frames or VHreplacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements," *Immunology*, vol. 119, no. 2, pp. 265–277, Oct. 2006.
[106] M. Barak, N. S. Zuckerman, H. Edelman, R. Unger, and R. Mehr, "IgTree©: Creating Immunoglobulin variable region gene lineage trees," *J. Immunol. Methods*, vol. 338, no. 1–2, pp. 67–74, Sep. 2008.
[107] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Genome Res.*, Jun. 2009.
[108] H. R. Hoogenboom, "Selecting and screening recombinant antibody libraries," *Nat. Biotechnol.*, vol. 23, no. 9, pp. 1105–1116, Sep. 2005.
[109] G. Köhler and C. Milstein, "Pillars Article: Continuous cultures of fused cells secreting antibody of predefined specificity. Nature, 1975, 256 (5517): 495–497.," *J. Immunol.*, vol. 174, no. 5, pp. 2453–2455, Mar. 2005.
[110] E. Traggiai, S. Becker, K. Subbarao, L. Kolesnikova, Y. Uematsu, M. R. Gismondo, B. R. Murphy, R. Rappuoli, and A. Lanzavecchia, "An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus," *Nat. Med.*, vol. 10, no. 8, pp. 871–875, Aug. 2004.
[111] M. J. Kwakkenbos, S. A. Diehl, E. Yasuda, A. Q. Bakker, C. M. M. van Geelen, M. V. Lukens, G. M. van Bleek, M. N. Widjojoatmodjo, W. M. J. M. Bogers, H. Mei, A. Radbruch, F. A. Scheeren, H. Spits, and T. Beaumont, "Generation of stable monoclonal antibody-producing B cell receptor-positive human memory B cells by genetic programming," *Nat Med*, vol. 16, no. 1, pp. 123–128, Jan. 2010.
[112] J. Wrammert, K. Smith, J. Miller, W. A. Langley, K. Kokko, C. Larsen, N.-Y. Zheng, I. Mays, L. Garman, C. Helms, J. James, G. M. Air, J. D. Capra, R. Ahmed, and P. C. Wilson, "Rapid cloning of high-affinity human monoclonal antibodies against influenza virus," *Nature*, vol. 453, no. 7195, pp. 667–671, May 2008.
[113] P.-J. Meijer, P. S. Andersen, M. Haahr Hansen, L. Steinaa, A. Jensen, J. Lantto, M. B. Oleksiewicz, K. Tengbjerg, T. R. Poulsen, V. W. Coljee, S. Bregenholt, J. S. Haurum, and L. S. Nielsen, "Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing," *J. Mol. Biol.*, vol. 358, no. 3, pp. 764–772, May 2006.

[114] T. Clackson, H. R. Hoogenboom, A. D. Griffiths, and G. Winter, "Making antibody fragments using phage display libraries," *Nature*, vol. 352, no. 6336, pp. 624–628, Aug. 1991.
[115] M. J. Feldhaus, R. W. Siegel, L. K. Opresko, J. R. Coleman, J. M. W. Feldhaus, Y. A. Yeung, J. R. Cochran, P. Heinzelman, D. Colby, J. Swers, C. Graff, H. S. Wiley, and K. D. Wittrup, "Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library," *Nat. Biotechnol.*, vol. 21, no. 2, pp. 163–170, Feb. 2003.
[116] B. R. Harvey, G. Georgiou, A. Hayhurst, K. J. Jeong, B. L. Iverson, and G. K. Rogers, "Anchored periplasmic expression, a versatile technology for the isolation of high-affinity antibodies from Escherichia coli-expressed libraries," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 25, pp. 9193–9198, Jun. 2004.
[117] C. Schaffitzel, J. Hanes, L. Jermutus, and A. Plückthun, "Ribosome display: an in vitro method for selection and evolution of antibodies from libraries," *J. Immunol. Methods*, vol. 231, no. 1–2, pp. 119–135, Dec. 1999.
[118] Y. Mazor, T. Van Blarcom, R. Mabry, B. L. Iverson, and G. Georgiou, "Isolation of engineered, full-length antibodies from libraries expressed in Escherichia coli," *Nat. Biotechnol.*, vol. 25, no. 5, pp. 563–565, May 2007.
[119] A. Jin, T. Ozawa, K. Tajiri, T. Obata, S. Kondo, K. Kinoshita, S. Kadowaki, K. Takahashi, T. Sugiyama, H. Kishi, and A. Muraguchi, "A rapid and efficient single-cell manipulation method for screening antigen-specific antibody-secreting cells from human peripheral blood," *Nat. Med.*, vol. 15, no. 9, pp. 1088–1092, Sep. 2009.
[120] J. C. Love, J. L. Ronan, G. M. Grotenbreg, A. G. van der Veen, and H. L. Ploegh, "A microengraving method for rapid selection of single cells producing antigen-specific antibodies," *Nat. Biotechnol.*, vol. 24, no. 6, pp. 703–707, Jun. 2006.
[121] C. W. Cobaugh, J. C. Almagro, M. Pogson, B. Iverson, and G. Georgiou, "Synthetic antibody libraries focused towards peptide ligands," *J. Mol. Biol.*, vol. 378, no. 3, pp. 622–633, May 2008.
[122] H. Persson, J. Lantto, and M. Ohlin, "A focused antibody library for improved hapten recognition," *J. Mol. Biol.*, vol. 357, no. 2, pp. 607–620, Mar. 2006.
[123] R. A. Manz, A. E. Hauser, F. Hiepe, and A. Radbruch, "Maintenance of Serum Antibody Levels," *Annu. Rev. Immunol.*, vol. 23, no. 1, pp. 367–386, 2005.
[124] M. Shapiro-Shelef and K. Calame, "Regulation of plasma-cell development," *Nat Rev Immunol*, vol. 5, no. 3, pp. 230–242, Mar. 2005.
[125] A. Krebber, S. Bornhauser, J. Burmester, A. Honegger, J. Willuda, H. R. Bosshard, and A. Plückthun, "Reliable cloning of functional antibody variable domains from hybridomas and spleen cell repertoires employing a reengineered phage display system," *J. Immunol. Methods*, vol. 201, no. 1, pp. 35–55, Feb. 1997.
[126] J. C. Cox, J. Lape, M. A. Sayed, and H. W. Hellinga, "Protein fabrication automation," *Protein Sci. Publ. Protein Soc.*, vol. 16, no. 3, pp. 379–390, Mar. 2007.

[127] A. Hayhurst, S. Happe, R. Mabry, Z. Koch, B. L. Iverson, and G. Georgiou, "Isolation and expression of recombinant antibody fragments to the biological warfare pathogen Brucella melitensis," *J. Immunol. Methods*, vol. 276, no. 1–2, pp. 185–196, May 2003.
[128] X. Gao, P. Yo, A. Keith, T. J. Ragan, and T. K. Harris, "Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences," *Nucleic Acids Res.*, vol. 31, no. 22, p. e143, Nov. 2003.
[129] Y. Mazor, I. Barnea, I. Keydar, and I. Benhar, "Antibody internalization studied using a novel IgG binding toxin fusion," *J. Immunol. Methods*, vol. 321, no. 1–2, pp. 41–59, Apr. 2007.
[130] J. A. Weinstein, N. Jiang, R. A. White, D. S. Fisher, and S. R. Quake, "High-Throughput Sequencing of the Zebrafish Antibody Repertoire," *Science*, vol. 324, no. 5928, pp. 807–810, May 2009.
[131] X. Ge, Y. Mazor, S. P. Hunicke-Smith, A. D. Ellington, and G. Georgiou, "Rapid construction and characterization of synthetic antibody libraries without DNA amplification," *Biotechnol. Bioeng.*, vol. 106, no. 3, pp. 347–357, Jun. 2010.
[132] T. G. Phan, D. Paus, T. D. Chan, M. L. Turner, S. L. Nutt, A. Basten, and R. Brink, "High affinity germinal center B cells are actively selected into the plasma cell compartment," *J. Exp. Med.*, vol. 203, no. 11, pp. 2419 –2424, Oct. 2006.
[133] R. Carlson, "The changing economics of DNA synthesis," *Nat. Biotechnol.*, vol. 27, no. 12, pp. 1091–1094, Dec. 2009.
[134] E. Lai, R. K. Wilson, and L. E. Hood, "Physical maps of the mouse and human immunoglobulin-like loci," *Adv. Immunol.*, vol. 46, pp. 1–59, 1989.
[135] K. Rajewsky, "Clonal selection and learning in the antibody system," *Nature*, vol. 381, no. 6585, pp. 751–758, Jun. 1996.
[136] I. Sanz, "Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions," *J. Immunol.*, vol. 147, no. 5, pp. 1720 –1729, 1991.
[137] H. Brezinschek, R. Brezinschek, and P. Lipsky, "Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction," *J. Immunol.*, vol. 155, no. 1, pp. 190 –202, Jul. 1995.
[138] J. Glanville, T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, and J. Pons, "Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation," *Proc. Natl. Acad. Sci.*, vol. 108, no. 50, pp. 20066 – 20071, Dec. 2011.
[139] R. L. Schelonka, J. Tanner, Y. Zhuang, G. L. Gartland, M. Zemlin, and H. W. Schroeder, "Categorical selection of the antibody repertoire in splenic B cells," *Eur. J. Immunol.*, vol. 37, no. 4, pp. 1010–1021, Apr. 2007.
[140] I. Suzuki, L. Pfister, A. Glas, C. Nottenburg, and E. C. Milner, "Representation of rearranged VH gene segments in the human adult antibody repertoire," *J. Immunol. Baltim. Md 1950*, vol. 154, no. 8, pp. 3902–3911, Apr. 1995.

[141] M. Yamada, R. Wasserman, B. A. Reichard, S. Shane, A. J. Caton, and G. Rovera, "Preferential utilization of specific immunoglobulin heavy chain diversity and joining segments in adult human peripheral blood B lymphocytes.," *J. Exp. Med.*, vol. 173, no. 2, pp. 395 –407, Feb. 1991.

[142] S. D. Boyd, B. A. Gaeta, K. J. Jackson, A. Z. Fire, E. L. Marshall, J. D. Merker, J. M. Maniar, L. N. Zhang, B. Sahaf, C. D. Jones, B. B. Simen, B. Hanczaruk, K. D. Nguyen, K. C. Nadeau, M. Egholm, D. B. Miklos, J. L. Zehnder, and A. M. Collins, "Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements," *J Immunol*, vol. 184, no. 12, pp. 6986–6992, Jun. 2010.

[143] J. P. Cox, I. M. Tomlinson, and G. Winter, "A directory of human germ-line V kappa segments reveals a strong bias in their usage," *Eur. J. Immunol.*, vol. 24, no. 4, pp. 827–836, Apr. 1994.

[144] S. J. Foster, H. P. Brezinschek, R. I. Brezinschek, and P. E. Lipsky, "Molecular mechanisms and selective influences that shape the kappa gene repertoire of IgM+ B cells," *J. Clin. Invest.*, vol. 99, no. 7, pp. 1614–1627, Apr. 1997.

[145] K. J. L. Jackson, Y. Wang, B. A. Gaeta, W. Pomat, P. Siba, J. Rimmer, W. A. Sewell, and A. M. Collins, "Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells," *Immunogenetics*, vol. 64, no. 1, pp. 3–14, Jul. 2011.

[146] O. Ignatovich, I. M. Tomlinson, P. T. Jones, and G. Winter, "The creation of diversity in the human immunoglobulin Vλ repertoire," *J. Mol. Biol.*, vol. 268, no. 1, pp. 69–77, Apr. 1997.

[147] S. L. Bridges, "Frequent N addition and clonal relatedness among immunoglobulin lambda light chains expressed in rheumatoid arthritis synovia and PBL, and the influence of V lambda gene segment utilization on CDR3 length.," *Mol. Med.*, vol. 4, no. 8, pp. 525–553, Aug. 1998.

[148] O. Ignatovich, I. M. Tomlinson, A. V. Popov, M. Brüggemann, and G. Winter, "Dominance of intrinsic genetic factors in shaping the human immunoglobulin Vλ repertoire," *J. Mol. Biol.*, vol. 294, no. 2, pp. 457–465, Nov. 1999.

[149] N. L. Farner, T. Dörner, and P. E. Lipsky, "Molecular Mechanisms and Selection Influence the Generation of the Human VλJλ Repertoire," *J. Immunol.*, vol. 162, no. 4, pp. 2137–2145, Feb. 1999.

[150] J. Lee, N. L. Monson, and P. E. Lipsky, "The VλJλ Repertoire in Human Fetal Spleen: Evidence for Positive Selection and Extensive Receptor Editing," *J. Immunol.*, vol. 165, no. 11, pp. 6322–6333, Dec. 2000.

[151] P. Richl, U. Stern, P. E. Lipsky, and H. J. Girschick, "The lambda gene immunoglobulin repertoire of human neonatal B cells," *Mol. Immunol.*, vol. 45, no. 2, pp. 320–327, Jan. 2008.

[152] K. Kawasaki, S. Minoshima, E. Nakato, K. Shibuya, A. Shintani, J. L. Schmeits, J. Wang, and N. Shimizu, "One-megabase sequence analysis of the human

immunoglobulin lambda gene locus," *Genome Res*., vol. 7, no. 3, pp. 250–261, Mar. 1997.
[153] K. Kawasaki, S. Minoshima, K. Schooler, J. Kudoh, S. Asakawa, P. J. de Jong, and N. Shimizu, "The organization of the human immunoglobulin lambda gene locus," *Genome Res*., vol. 5, no. 2, pp. 125–135, Sep. 1995.
[154] S. C. Williams, J.-P. Frippiat, I. M. Tomlinson, O. Ignatovich, M.-P. Lefranc, and G. Winter, "Sequence and Evolution of the Human Germline VλRepertoire," *J. Mol. Biol*., vol. 264, no. 2, pp. 220–232, Nov. 1996.
[155] R. Saada, M. Weinberger, G. Shahaf, and R. Mehr, "Models for antigen receptor gene rearrangement: CDR3 length," *Immunol Cell Biol*, vol. 85, no. 4, pp. 323–332, Apr. 2007.
[156] E. A. Padlan, "Anatomy of the antibody molecule," *Mol. Immunol*., vol. 31, no. 3, pp. 169–217, Feb. 1994.
[157] I. A. Wilson and R. L. Stanfield, "Antibody-antigen interactions: new structures and new conformational changes," *Curr. Opin. Struct. Biol*., vol. 4, no. 6, pp. 857–867, Dec. 1994.
[158] D. Kuroda, H. Shirai, M. Kobori, and H. Nakamura, "Systematic classification of CDR-L3 in antibodies: Implications of the light chain subtypes and the VL–VH interface," *Proteins Struct. Funct. Bioinforma*., vol. 75, no. 1, pp. 139–146, 2009.
[159] N. Schoettler, D. Ni, and M. Weigert, "B cell receptor light chain repertoires show signs of selection with differences between groups of healthy individuals and SLE patients," *Mol. Immunol*., vol. 51, no. 3–4, pp. 273–282, Jul. 2012.
[160] D. K. Lanning, K.-J. Rhee, and K. L. Knight, "Intestinal bacteria and development of the B-lymphocyte repertoire," *Trends Immunol*., vol. 26, no. 8, pp. 419–425, Aug. 2005.
[161] K. L. Knight, "Restricted VH Gene Usage and Generation of Antibody Diversity in Rabbit," *Annu. Rev. Immunol*., vol. 10, no. 1, pp. 593–616, 1992.
[162] S. Dray, S. Dubiski, A. Kelus, E. S. Lennox, and J. Oudin, "A notation for allotypy," *Nature*, vol. 195, pp. 785–786, Aug. 1962.
[163] B. S. Kim and S. Dray, "Expression of the a, x, and y variable region genes of heavy chains among IgG, IgM, and IgA molecules of normal and a locus allotype-suppressed rabbits," *J. Immunol. Baltim. Md 1950*, vol. 111, no. 3, pp. 750–760, Sep. 1973.
[164] S. Dray, G. O. Young, and A. Nisonoff, "DISTRIBUTION OF ALLOTYPIC SPECIFICITIES AMONG RABBIT GAMMA-GLOBULIN MOLECULES GENETICALLY DEFINED AT TWO LOCI," *Nature*, vol. 199, pp. 52–55, Jul. 1963.
[165] R. G. Mage, D. Lanning, and K. L. Knight, "B cell and antibody repertoire development in rabbits: the requirement of gut-associated lymphoid tissues," *Dev. Comp. Immunol*., vol. 30, no. 1–2, pp. 137–153, 2006.
[166] E. M. Gertz, A. A. Schäffer, R. Agarwala, A. Bonnet-Garnier, C. Rogel-Gaillard, H. Hayes, and R. G. Mage, "Accuracy and coverage assessment of Oryctolagus

cuniculus (rabbit) genes encoding immunoglobulins in the whole genome sequence assembly (OryCun2.0) and localization of the IGH locus to chromosome 20," *Immunogenetics*, vol. 65, no. 10, pp. 749–762, Oct. 2013.

[167] K. H. Roux, P. Dhanarajan, V. Gottschalk, W. T. McCormick, and R. W. Renshaw, "Latent a1 VH germline genes in an a2a2 rabbit. Evidence for gene conversion at both the germline and somatic levels," *J. Immunol. Baltim. Md 1950*, vol. 146, no. 6, pp. 2027–2036, Mar. 1991.

[168] P. A. Larsen and T. P. L. Smith, "Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire," *BMC Immunol.*, vol. 13, p. 52, 2012.

[169] F. Wang, D. C. Ekiert, I. Ahmad, W. Yu, Y. Zhang, O. Bazirgan, A. Torkamani, T. Raudsepp, W. Mwangi, M. F. Criscitiello, I. A. Wilson, P. G. Schultz, and V. V. Smider, "Reshaping antibody diversity," *Cell*, vol. 153, no. 6, pp. 1379–1393, Jun. 2013.

[170] D. Sehgal, G. Johnson, T. T. Wu, and R. G. Mage, "Generation of the primary antibody repertoire in rabbits: expression of a diverse set of Igk-V genes may compensate for limited combinatorial diversity at the heavy chain locus," *Immunogenetics*, vol. 50, no. 1–2, pp. 31–42, Oct. 1999.

[171] P. D. Weinstein, A. O. Anderson, and R. G. Mage, "Rabbit IgH sequences in appendix germinal centers: VH diversification by gene conversion-like and hypermutation mechanisms," *Immunity*, vol. 1, no. 8, pp. 647–659, Nov. 1994.

[172] R. S. Becker and K. L. Knight, "Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits," *Cell*, vol. 63, no. 5, pp. 987–997, Nov. 1990.

[173] D. Lanning, P. Sethupathi, K. J. Rhee, S. K. Zhai, and K. L. Knight, "Intestinal microflora and diversification of the rabbit antibody repertoire," *J. Immunol. Baltim. Md 1950*, vol. 165, no. 4, pp. 2012–2019, Aug. 2000.

[174] R. S. Harris, J. E. Sale, S. K. Petersen-Mahrt, and M. S. Neuberger, "AID is essential for immunoglobulin V gene conversion in a cultured B cell line," *Curr. Biol. CB*, vol. 12, no. 5, pp. 435–438, Mar. 2002.

[175] H. Arakawa and J.-M. Buerstedde, "Immunoglobulin gene conversion: insights from bursal B cells and the DT40 cell line," *Dev. Dyn. Off. Publ. Am. Assoc. Anat.*, vol. 229, no. 3, pp. 458–464, Mar. 2004.

[176] C. A. Reynaud, A. Dahan, V. Anquez, and J. C. Weill, "Somatic hyperconversion diversifies the single Vh gene of the chicken with a high incidence in the D region," *Cell*, vol. 59, no. 1, pp. 171–183, Oct. 1989.

[177] W. J. J. Finlay, L. Bloom, and O. Cunningham, "Optimized generation of high-affinity, high-specificity single-chain Fv antibodies from multiantigen immunized chickens," *Methods Mol. Biol. Clifton NJ*, vol. 681, pp. 383–401, 2011.

[178] F. Ros, J. Puels, N. Reichenberger, W. van Schooten, R. Buelow, and J. Platzer, "Sequence analysis of 0.5 Mb of the rabbit germline immunoglobulin heavy chain locus," *Gene*, vol. 330, pp. 49–59, Apr. 2004.

[179] X. Zhu, A. Boonthum, S. K. Zhai, and K. L. Knight, "B lymphocyte selection and age-related changes in VH gene usage in mutant Alicia rabbits," *J. Immunol. Baltim. Md 1950*, vol. 163, no. 6, pp. 3313–3320, Sep. 1999.
[180] K. E. Bernstein, C. B. Alexander, and R. G. Mage, "Germline VH genes in an a3 rabbit not typical of any one VHa allotype," *J. Immunol. Baltim. Md 1950*, vol. 134, no. 5, pp. 3480–3488, May 1985.
[181] M. G. Fitts and D. W. Metzger, "Identification of rabbit genomic Ig-VH pseudogenes that could serve as donor sequences for latent allotype expression," *J. Immunol. Baltim. Md 1950*, vol. 145, no. 8, pp. 2713–2717, Oct. 1990.
[182] K. L. Knight and R. S. Becker, "Molecular basis of the allelic inheritance of rabbit immunoglobulin VH allotypes: implications for the generation of antibody diversity," *Cell*, vol. 60, no. 6, pp. 963–970, Mar. 1990.
[183] C. Raman, H. Spieker-Polet, P. C. Yam, and K. L. Knight, "Preferential VH gene usage in rabbit Ig-secreting heterohybridomas," *J. Immunol. Baltim. Md 1950*, vol. 152, no. 8, pp. 3935–3945, Apr. 1994.
[184] M. L. Friedman, C. Tunyaplin, S. K. Zhai, and K. L. Knight, "Neonatal VH, D, and JH gene usage in rabbit B lineage cells," *J. Immunol. Baltim. Md 1950*, vol. 152, no. 2, pp. 632–641, Jan. 1994.
[185] R. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, no. 1, p. 113, 2004.
[186] J. Pelé, J.-M. Bécu, H. Abdi, and M. Chabbert, "Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling," *BMC Bioinformatics*, vol. 13, p. 133, 2012.
[187] S. Sawyer, "Statistical tests for detecting gene conversion," *Mol. Biol. Evol.*, vol. 6, no. 5, pp. 526–538, Sep. 1989.
[188] W. S. Torgerson, "Multidimensional scaling of similarity," *Psychometrika*, vol. 30, no. 4, pp. 379–393, Dec. 1965.
[189] J. Pelé, H. Abdi, M. Moreau, D. Thybert, and M. Chabbert, "Multidimensional scaling reveals the main evolutionary pathways of class A G-protein-coupled receptors," *PloS One*, vol. 6, no. 4, p. e19094, 2011.
[190] D. G. Higgins, "Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets," *Comput. Appl. Biosci. CABIOS*, vol. 8, no. 1, pp. 15–22, Feb. 1992.
[191] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.
[192] K.-J. Rhee, P. J. Jasper, P. Sethupathi, M. Shanmugam, D. Lanning, and K. L. Knight, "Positive selection of the peripheral B cell repertoire in gut-associated lymphoid tissues," *J. Exp. Med.*, vol. 201, no. 1, pp. 55–62, Jan. 2005.
[193] P. J. Esteves, D. Lanning, N. Ferrand, K. L. Knight, S. K. Zhai, and W. van der Loo, "The evolution of the immunoglobulin heavy chain variable region (IgVH) in Leporids: an unusual case of transspecies polymorphism," *Immunogenetics*, vol. 57, no. 11, pp. 874–882, Dec. 2005.

[194] E. Appella, A. Chersi, J. Rejnek, R. Reisfeld, and R. Mage, "Rabbit immunoglobulin lambda chains: isolation and amino acid sequence of cysteine-containing peptides," *Immunochemistry*, vol. 11, no. 8, pp. 395–402, Aug. 1974.
[195] S. Dubiski and P. J. Muller, "A 'new' allotypic specificity (A9) of rabbit immunoglobulin," *Nature*, vol. 214, no. 5089, pp. 696–697, May 1967.
[196] H. W. Schroeder, G. C. Ippolito, and S. Shiokawa, "Regulation of the antibody repertoire through control of HCDR3 diversity," *Vaccine*, vol. 16, no. 14–15, pp. 1383–1390, Sep. 1998.
[197] L. Wu, K. Oficjalska, M. Lambert, B. J. Fennell, A. Darmanin-Sheehan, D. Ní Shúilleabháin, B. Autin, E. Cummins, L. Tchistiakova, L. Bloom, J. Paulsen, D. Gill, O. Cunningham, and W. J. J. Finlay, "Fundamental characteristics of the immunoglobulin VH repertoire of chickens in comparison with those of humans, mice, and camelids," *J. Immunol. Baltim. Md 1950*, vol. 188, no. 1, pp. 322–333, Jan. 2012.
[198] A. Gilles, E. Meglécz, N. Pech, S. Ferreira, T. Malausa, and J.-F. Martin, "Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing," *BMC Genomics*, vol. 12, p. 245, 2011.
[199] J. M. Darlow and D. I. Stott, "Gene conversion in human rearranged immunoglobulin genes," *Immunogenetics*, vol. 58, no. 7, pp. 511–522, Jul. 2006.
[200] N. D'Avirro, D. Truong, B. Xu, and E. Selsing, "Sequence transfers between variable regions in a mouse antibody transgene can occur by gene conversion," *J. Immunol. Baltim. Md 1950*, vol. 175, no. 12, pp. 8133–8137, Dec. 2005.
[201] B. Duvvuri and G. E. Wu, "Gene Conversion-Like Events in the Diversification of Human Rearranged IGHV3-23*01 Gene Sequences," *Front. Immunol.*, vol. 3, p. 158, 2012.
[202] J. S. Huston, "Engineering antibodies for the 21st century," *Protein Eng. Des. Sel. PEDS*, vol. 25, no. 10, pp. 483–484, Oct. 2012.
[203] R. Castro, L. Jouneau, H.-P. Pham, O. Bouchez, V. Giudicelli, M.-P. Lefranc, E. Quillet, A. Benmansour, F. Cazals, A. Six, S. Fillatreau, O. Sunyer, and P. Boudinot, "Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection," *PLoS Pathog.*, vol. 9, no. 1, p. e1003098, Jan. 2013.
[204] H. Arakawa, K. Kuma, M. Yasuda, S. Ekino, A. Shimizu, and H. Yamagishi, "Effect of environmental antigens on the Ig diversification and the selection of productive V-J joints in the bursa," *J. Immunol. Baltim. Md 1950*, vol. 169, no. 2, pp. 818–828, Jul. 2002.
[205] A. C. Logan, H. Gao, C. Wang, B. Sahaf, C. D. Jones, E. L. Marshall, I. Buño, R. Armstrong, A. Z. Fire, K. I. Weinberg, M. Mindrinos, J. L. Zehnder, S. D. Boyd, W. Xiao, R. W. Davis, and D. B. Miklos, "High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 52, pp. 21194–21199, Dec. 2011.

[206] N. Hosen, Y. Matsuoka, S. Kishida, J. Nakata, Y. Mizutani, K. Hasegawa, A. Mugitani, H. Ichihara, Y. Aoyama, S. Nishida, A. Tsuboi, F. Fujiki, N. Tatsumi, H. Nakajima, M. Hino, T. Kimura, K. Yata, M. Abe, Y. Oka, Y. Oji, A. Kumanogoh, and H. Sugiyama, "CD138-negative clonogenic cells are plasma cells but not B cells in some multiple myeloma patients," *Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund UK*, vol. 26, no. 9, pp. 2135–2141, Sep. 2012.

[207] "Immunoglobulin lambda isotype gene rearrangements in B cell malignancies," *Publ. Online 18 January 2001 Doi101038sjleu2401985*, vol. 15, no. 1, Jan. 2001.

[208] P. Parameswaran, Y. Liu, K. M. Roskin, K. K. L. Jackson, V. P. Dixit, J.-Y. Lee, K. L. Artiles, S. Zompi, M. J. Vargas, B. B. Simen, B. Hanczaruk, K. R. McGowan, M. A. Tariq, N. Pourmand, D. Koller, A. Balmaseda, S. D. Boyd, E. Harris, and A. Z. Fire, "Convergent Antibody Signatures in Human Dengue," *Cell Host Microbe*, vol. 13, no. 6, pp. 691–700, Jun. 2013.

[209] K. J. L. Jackson, Y. Liu, K. M. Roskin, J. Glanville, R. A. Hoh, K. Seo, E. L. Marshall, T. C. Gurley, M. A. Moody, B. F. Haynes, E. B. Walter, H.-X. Liao, R. A. Albrecht, A. García-Sastre, J. Chaparro-Riggers, A. Rajpal, J. Pons, B. B. Simen, B. Hanczaruk, C. L. Dekker, J. Laserson, D. Koller, M. M. Davis, A. Z. Fire, and S. D. Boyd, "Human Responses to Influenza Vaccination Show Seroconversion Signatures and Convergent Antibody Rearrangements," *Cell Host Microbe*, vol. 16, no. 1, pp. 105–114, Jul. 2014.

[210] E. S. Mroczek, G. C. Ippolito, T. Rogosch, K. H. Hoi, T. A. Hwangpo, M. G. Brand, Y. Zhuang, C. R. Liu, D. A. Schneider, M. Zemlin, E. E. Brown, G. Georgiou, and H. W. J. Schroeder, "Differences in the composition of the human antibody repertoire by B cell subsets in the blood," *B Cell Biol.*, vol. 5, p. 96, 2014.

[211] K. Kedzierska, N. L. La Gruta, J. Stambas, S. J. Turner, and P. C. Doherty, "Tracking phenotypically and functionally distinct T cell subsets via T cell repertoire diversity," *Mol. Immunol.*, vol. 45, no. 3, pp. 607–618, Feb. 2008.

[212] V. Venturi, M. F. Quigley, H. Y. Greenaway, P. C. Ng, Z. S. Ende, T. McIntosh, T. E. Asher, J. R. Almeida, S. Levy, D. A. Price, M. P. Davenport, and D. C. Douek, "A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing," *J. Immunol. Baltim. Md 1950*, vol. 186, no. 7, pp. 4285–4294, Apr. 2011.

[213] G. Kopsidas, R. K. Carman, E. L. Stutt, A. Raicevic, A. S. Roberts, M.-A. V. Siomos, N. Dobric, L. Pontes-Braz, and G. Coia, "RNA mutagenesis yields highly diverse mRNA libraries for in vitro protein evolution," *BMC Biotechnol.*, vol. 7, no. 1, p. 18, Apr. 2007.

[214] J. D. Roberts, K. Bebenek, and T. A. Kunkel, "The accuracy of reverse transcriptase from HIV-1," *Science*, vol. 242, no. 4882, pp. 1171–1173, Nov. 1988.

[215] P. McInerney, P. Adams, and M. Z. Hadi, "Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase," *Mol. Biol. Int.*, vol. 2014, p. e287430, Aug. 2014.

[216] P. Nguyen, J. Ma, D. Pei, C. Obert, C. Cheng, and T. L. Geiger, "Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire," *BMC Genomics*, vol. 12, no. 1, p. 106, Feb. 2011.
[217] D. A. Bolotin, I. Z. Mamedov, O. V. Britanova, I. V. Zvyagin, D. Shagin, S. V. Ustyugova, M. A. Turchaninova, S. Lukyanov, Y. B. Lebedev, and D. M. Chudakov, "Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms," *Eur. J. Immunol.*, vol. 42, no. 11, pp. 3073–3083, Nov. 2012.
[218] A. Ratan, W. Miller, J. Guillory, J. Stinson, S. Seshagiri, and S. C. Schuster, "Comparison of sequencing platforms for single nucleotide variant calls in a human sample," *PloS One*, vol. 8, no. 2, p. e55089, 2013.
[219] K. Robasky, N. E. Lewis, and G. M. Church, "The role of replicates for error mitigation in next-generation sequencing," *Nat. Rev. Genet.*, vol. 15, no. 1, pp. 56–62, Jan. 2014.
[220] I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein, "Detection and quantification of rare mutations with massively parallel sequencing," *Proc. Natl. Acad. Sci.*, vol. 108, no. 23, pp. 9530 –9535, Jun. 2011.
[221] M. W. Schmitt, S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, and L. A. Loeb, "Detection of ultra-rare mutations by next-generation sequencing," *Proc. Natl. Acad. Sci.*, vol. 109, no. 36, pp. 14508–14513, Sep. 2012.
[222] C. Vollmers, R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake, "Genetic measurement of memory B-cell recall using antibody repertoire sequencing," *Proc. Natl. Acad. Sci.*, Jul. 2013.
[223] M. Shugay, O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, S. Lukyanov, T. N. Schumacher, and D. M. Chudakov, "Towards error-free profiling of immune repertoires," *Nat. Methods*, vol. advance online publication, May 2014.
[224] D. I. Lou, J. A. Hussmann, R. M. McBee, A. Acevedo, R. Andino, W. H. Press, and S. L. Sawyer, "High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing," *Proc. Natl. Acad. Sci.*, p. 201319590, Nov. 2013.
[225] Y. Mazor, T. Van Blarcom, B. L. Iverson, and G. Georgiou, "E-clonal antibodies: selection of full-length IgG antibodies using bacterial periplasmic display," *Nat. Protoc.*, vol. 3, no. 11, pp. 1766–1777, Oct. 2008.
[226] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, Jul. 2011.

[227] B. Arezi and H. Hogrefe, "Novel mutations in Moloney Murine Leukemia Virus reverse transcriptase increase thermostability through tighter binding to template-primer," *Nucleic Acids Res.*, vol. 37, no. 2, pp. 473–481, Feb. 2009.
[228] H. Ochman, A. S. Gerber, and D. L. Hartl, "Genetic applications of an inverse polymerase chain reaction," *Genetics*, vol. 120, no. 3, pp. 621–623, Nov. 1988.
[229] V. J. Gadkar and M. Filion, "A novel method to perform genomic walks using a combination of single strand DNA circularization and rolling circle amplification," *J. Microbiol. Methods*, vol. 87, no. 1, pp. 38–43, Oct. 2011.
[230] D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, "Analysis of membrane and surface protein sequences with the hydrophobic moment plot," *J. Mol. Biol.*, vol. 179, no. 1, pp. 125–142, Oct. 1984.
[231] U. Ravn, F. Gueneau, L. Baerlocher, M. Osteras, M. Desmurs, P. Malinge, G. Magistrelli, L. Farinelli, M. H. Kosco-Vilbois, and N. Fischer, "By-passing in vitro screening--next generation sequencing technologies applied to antibody display and in silico candidate selection," *Nucleic Acids Res.*, vol. 38, no. 21, p. e193, Nov. 2010.